Journal of Philosophical Logic (2021) 50:1341–1374
https://doi.org/10.1007/s10992-021-09601-z

Causal Sufficiency and Actual Causation

Sander Beckers[1]

Received: 30 June 2020 / Accepted: 23 March 2021 / Published online: 19 June 2021

Abstract
Pearl opened the door to formally defining *actual causation* using causal models. His approach rests on two strategies: first, capturing the widespread intuition that $X = x$ causes $Y = y$ iff $X = x$ is a *Necessary Element of a Sufficient Set* for $Y = y$, and second, showing that his definition gives intuitive answers on a wide set of problem cases. This inspired dozens of variations of his definition of actual causation, the most prominent of which are due to Halpern & Pearl. Yet all of them ignore Pearl's first strategy, and the second strategy taken by itself is unable to deliver a consensus. This paper offers a way out by going back to the first strategy: it offers six formal definitions of *causal sufficiency* and two interpretations of necessity. Combining the two gives twelve new definitions of actual causation. Several interesting results about these definitions and their relation to the various Halpern & Pearl definitions are presented. Afterwards the second strategy is evaluated as well. In order to maximize neutrality, the paper relies mostly on the examples and intuitions of Halpern & Pearl. One definition comes out as being superior to all others, and is therefore suggested as a new definition of actual causation.

Keywords Actual causation · Causal sufficiency · NESS · Counterfactuals

1 Introduction

Two decades have passed since Judea Pearl's groundbreaking book on causality was published [16]. It offers a formal account of causal models that led causal modeling to become a central part of Artificial Intelligence. One of the book's most important applications for philosophy is its formal definition of *actual causation*, i.e., causation of particular events.

Pearl defends his account of actual causation using two strategies. The first strategy starts with the widely shared intuition that $X = x$ causes $Y = y$ iff $X = x$

✉ Sander Beckers
srekcebrednas@gmail.com

[1] Munich Center for Mathematical Philosophy, LMU Munich, Münich, Germany

is a *Necessary Element of a Sufficient Set* for $Y = y$ (the NESS intuition, from now on).[1,2] Pearl claims that using causal models allows one to make this intuition formally precise, whereas existing logical notions of necessity and sufficiency lack the resources to do so. The second strategy is to demonstrate that his formal account offers intuitive verdicts for a number of problematic examples.

Ever since, Pearl's account has come under severe criticism. By now there are dozens of papers – both from philosophers and from researchers in AI – attempting to improve upon his account.[3] Most prominently, Pearl himself has offered several revisions of his account in collaboration with Halpern, culminating in the most recent revision by Halpern individually [7–10, 17]. Together these accounts of causation are referred to as the Halpern & Pearl definitions, or *HP definitions* for short, and they are by far the most influential accounts of causation out there.

The problem with all of these attempts at revising Pearl's initial account, is that they completely ignore the first strategy and focus almost excusively on the second strategy. Roughly put, the typical setup is to go over some examples for which existing definitions give counterintuitive answers, and then to construct a new definition that does not do so. It is unrealistic to expect that this second strategy in and of itself can deliver a satisfactory account of causation, because there are too many examples and even more intuitions [3, 4].

To solve this problem, this paper starts out with an explicit focus on the first strategy. It is striking that immediately after discussing the NESS intuition, Pearl diverges into complicated technical notions like "sustenance" and "causal beams" and never looks back, be it in his book or in the subsequent work on the HP definitions. Instead I offer what is the most natural route down the first strategy, namely to look at formalizations of *causal sufficiency* (as opposed to logical sufficiency) and combine them with two interpretations of *necessity*. Taken together this results in twelve distinct formal definitions of actual causation.

These definitions are compared to each other and to the HP definitions, leading to several interesting results. For one, it turns out that one of these twelve definitions is equivalent to the most recent HP definition [7, 8]. Therefore this paper is the first to show that one of the HP definitions succeeds in delivering Pearl's promise. At the same time, it also shows that the other HP definitions do not.

Next we turn to the second strategy. Given the diversity of intuitions about the many examples presented in the literature, the best we can do is arrive at a comparative verdict: does one of the definitions here developed fare better than the HP definitions? In order to avoid relying on my own intuitions, I present two criteria by

[1] This acronym was coined by Wright [22], but Pearl does not intend to formalize the specific manner in which Wright understood it, nor do I in the current paper. I have formalized Wright's interpretation of the NESS definition elsewhere, in the process of developing another definition of causation [1]. The latter definition is in many ways a simplification of the definition that I defend here. The precise relation between these two definitions is the subject of future work.

[2] Mackie [13] formulates the same intuition differently, resulting in the equally famous INUS acronym. See Wright [23] for a detailed discussion of the subtle differences between them.

[3] Just to name some of the most influential ones: Hall [6], Hitchcock [11, 12], Weslake [20] and Woodward [21].

which we can answer this question. First, I make use of Halpern and Pearl's own examples and rely almost exclusively on their intuitions, which for the most part align with the consensus in the literature. (Example 6 forms a notable exception that was suggested to me by a reviewer.) Here the answer is that one of the twelve definitions does better than the HP definitions. Second, I present six examples that are very similar to each other, and assess which definitions are able to handle them in a consistent (and preferably also intuitive) manner. Here the answer is that the previous definition again does better than the HP definitions.

Therefore I suggest adopting this definition of actual causation. Roughly, this definition states that $X = x$ causes $Y = y$ iff there is a set $\mathbf{W} = \mathbf{w}$ so that $(X = x, \mathbf{W} = \mathbf{w})$ is sufficient for $Y = y$ along a causal network $\mathbf{N}$ *and* there exists some value x' so that $(X = x', \mathbf{W} = \mathbf{w})$ is not sufficient for $Y = y$ along any causal subnetwork of $\mathbf{N}$.

This paper is laid out as follows. The next section introduces *structural equations models*, the formal causal models that are used to express all the definitions. Then I state the three most recent HP definitions in Section 3. Section 4 presents six notions of causal sufficiency and shows how they relate to each other. We then use these six notions to formalize actual causation along the NESS intuition in Section 5, and discuss several interesting results. After this theoretical groundwork, we start looking for the best definition. Two definitions are discarded by showing that they have certain unacceptable properties in Section 6. Finally, Section 7 compares the remaining definitions to each other and to the HP definitions by considering examples from Halpern & Pearl and a few additional ones.

2 Structural Equations Modeling

This section reviews the definition of causal models as they were introduced by Pearl [16]. Much of the discussion and notation is taken from Halpern [8] with little change.

Definition 1 A signature $\mathcal{S}$ is a tuple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$, where $\mathcal{U}$ is a set of *exogenous* variables, $\mathcal{V}$ is a set of *endogenous* variables, and $\mathcal{R}$ a function that associates with every variable $Y \in \mathcal{U} \cup \mathcal{V}$ a nonempty set $\mathcal{R}(Y)$ of possible values for Y (i.e., the set of values over which Y *ranges*). If $\mathbf{X} = (X_1, \ldots, X_n)$, $\mathcal{R}(\mathbf{X})$ denotes the crossproduct $\mathcal{R}(X_1) \times \cdots \times \mathcal{R}(X_n)$.

Exogenous variables represent factors whose causal origins are outside the scope of the causal model, such as background conditions and noise. The values of the endogenous variables, on the other hand, are causally determined by other variables within the model (both endogenous and exogenous).

Definition 2 A *causal model* M is a pair $(\mathcal{S}, \mathcal{F})$, where $\mathcal{S}$ is a signature and $\mathcal{F}$ defines a function that associates with each endogenous variable X a *structural equation* F_X giving the value of X in terms of the values of other endogenous and exogenous variables. Formally, the equation F_X maps $\mathcal{R}(\mathcal{U} \cup \mathcal{V} - \{X\})$ to $\mathcal{R}(X)$, so F_X determines the value of X, given the values of all the other variables in $\mathcal{U} \cup \mathcal{V}$.

Note that there are no functions associated with exogenous variables; their values are determined outside the model. We call a setting $\mathbf{u} \in \mathcal{R}(\mathcal{U})$ of values of exogenous variables a *context*.

The value of X may depend on the values of only a few other variables. X *depends on* Y *in context* $\mathbf{u}$ if there is some setting of the endogenous variables other than X and Y such that if the exogenous variables have value $\mathbf{u}$, then varying the value of Y in that context results in a variation in the value of X; that is, there is a setting $\mathbf{z}$ of the endogenous variables other than X and Y and values y and y' of Y such that $F_X(y, \mathbf{z}, \mathbf{u}) \neq F_X(y', \mathbf{z}, \mathbf{u})$. We then say that Y is a *parent* of X.

We extend this genealogical terminology in the usual manner, by taking the *ancestor* relation to be the transitive closure of the parent relation (i.e., Y is an ancestor of X iff there exist variables so that Y is a parent of V_1, V_1 is a parent of V_2, ..., and V_n is a parent of X). The *descendant* relation is simply the reversal of the ancestor relation (i.e., X is a descendant of Y iff Y is an ancestor of X.) A *path* is a sequence of variables in which each element is a child of the previous element.

In this paper we restrict attention to *strongly recursive* (or *strongly acyclic*) models, that is, models where there is a partial order $\preceq$ on variables such that if Y depends on X, then $X \prec Y$. In a strongly recursive model, given a context $\mathbf{u}$, the values of all the remaining variables are determined (we can just solve for the value of the variables in the order given by $\preceq$). We often write the equation for an endogenous variable as $X = f(\mathbf{Y})$; this denotes that the value of X depends only on the values of the variables in $\mathbf{Y}$, and the connection is given by the function f. For example, we might have $X = Y + 5$.

An *intervention* has the form $\mathbf{X} \leftarrow \mathbf{x}$, where $\mathbf{X}$ is a set of endogenous variables. Intuitively, this means that the values of the variables in $\mathbf{X}$ are set to the values $\mathbf{x}$. The structural equations define what happens in the presence of interventions. Setting the value of some variables $\mathbf{X}$ to $\mathbf{x}$ in a causal model $M = (\mathcal{S}, \mathcal{F})$ results in a new causal model, denoted $M_{\mathbf{X} \leftarrow \mathbf{x}}$, which is identical to M, except that $\mathcal{F}$ is replaced by $\mathcal{F}^{\mathbf{X} \leftarrow \mathbf{x}}$: for each variable $Y \notin \mathbf{X}$, $F_Y^{\mathbf{X} \leftarrow \mathbf{x}} = F_Y$ (i.e., the equation for Y is unchanged), while for each X' in $\mathbf{X}$, the equation $F_{X'}$ for X' is replaced by $X' = x'$ (where x' is the value in $\mathbf{x}$ corresponding to X').

Given a signature $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$, an *atomic formula* is a formula of the form $X = x$, for $X \in \mathcal{V}$ and $x \in \mathcal{R}(X)$. A *causal formula (over* $\mathcal{S}$*)* is one of the form $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\phi$, where

- ϕ is a Boolean combination of atomic formulas,
- $Y_1, \ldots, Y_k$ are distinct variables in $\mathcal{V}$, and
- $y_i \in \mathcal{R}(Y_i)$ for each $1 \leq i \leq k$.

Such a formula is abbreviated as $[\mathbf{Y} \leftarrow \mathbf{y}]\phi$. The special case where $k = 0$ is abbreviated as ϕ. Intuitively, $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\phi$ says that ϕ would hold if Y_i were set to y_i, for $i = 1, \ldots, k$.

A causal formula ψ is true or false in a *causal setting*, which is a causal model given a context. As usual, we write $(M, \mathbf{u}) \models \psi$ if the causal formula ψ is true in the causal setting $(M, \mathbf{u})$. The $\models$ relation is defined inductively. $(M, \mathbf{u}) \models X = x$ if the variable X has value x in the unique (since we are dealing with recursive models) solution to the equations in M in context $\mathbf{u}$ (i.e., the unique vector of values that

simultaneously satisfies all equations in M with the variables in $\mathcal{U}$ set to $\mathbf{u}$). The truth of conjunctions and negations is defined in the standard way. Finally, $(M, \mathbf{u}) \models [\mathbf{Y} \leftarrow \mathbf{y}]\phi$ if $(M_{\mathbf{Y} \leftarrow \mathbf{y}}, \mathbf{u}) \models \phi$ (i.e., the intervention $\mathbf{Y} \leftarrow \mathbf{y}$ transforms M into a new model $M_{\mathbf{Y} \leftarrow \mathbf{y}}$, in which we assess the truth of ϕ).

3 HP Definitions

Now on to the HP definitions. As Pearl [16]'s initial definition is a precursor to the HP definitions that gives less intuitive results and is far more complicated, I do not discuss it. (It is safe to say that by now it has been unanimously rejected.) Two of the HP definitions are developed by both Halpern and Pearl, whereas the third one is solely due to Halpern. The relations between them are extensively discussed by Halpern [8].

The general form of all three definitions is as follows (where ϕ is a Boolean combination of atomic formulas):

Definition 3 $\mathbf{X} = \mathbf{x}$ is an *actual cause* of ϕ in $(M, \mathbf{u})$ if the following three conditions hold:

AC1. $(M, \mathbf{u}) \models (\mathbf{X} = \mathbf{x}) \wedge \phi$.
AC2. See below.
AC3. $\mathbf{X}$ is minimal; there is no strict subset $\mathbf{X}''$ of $\mathbf{X}$ such that $\mathbf{X}'' = \mathbf{x}''$ satisfies AC2, where $\mathbf{x}''$ is the restriction of $\mathbf{x}$ to the variables in $\mathbf{X}''$.

Questions of actual causation are posed relative to an *actual context* $\mathbf{u}$, because as we know from the previous section a context completely determines which events actually took place. So AC1 represents the trivial requirement that the candidate cause and effect are among the events which took place. AC3 is also fairly straightforward: we should not consider redundant elements to be parts of causes. The real content of the definition lies with AC2.

Throughout the rest of the paper, settings of variables $\mathbf{V}$ with superscript * (i.e., $\mathbf{v}^*$) indicate that $(M, \mathbf{u}) \models (\mathbf{V} = \mathbf{v}^*)$. Settings of variables $\mathbf{V}$ with superscript $'$ (i.e., $\mathbf{v}'$) indicate that $(M, \mathbf{u}) \models (V \neq v')$ for each $V \in \mathbf{V}$. Settings of variables without any superscript can refer to any setting.

In line with the NESS intuition, we should expect AC2 to consist of formal variants of these two conditions:[4]

AC2(b). There is a set $\mathbf{W}$ so that $(\mathbf{X} = \mathbf{x}, \mathbf{W} = \mathbf{w}^*)$ is causally sufficient for ϕ.
AC2(a). $\mathbf{X} = \mathbf{x}$ is necessary for the sufficiency of $(\mathbf{X} = \mathbf{x}, \mathbf{W} = \mathbf{w}^*)$.

At first glance, the first two HP definitions seem to meet this expectation: they consist of conditions AC2(a) and AC2(b), and Halpern refers to these as a "necessity condition" and a "sufficiency condition" [7, p. 3]. Upon closer examination, however,

[4] I list them unalphabetically for consistency with the HP definitions.

it is hard to see how either version of AC2(b) can sensibly be interpreted as capturing causal sufficiency.

We start with **Original HP** [9]:

Definition 4 [**Original HP**]

AC2(a). There is a partition of $\mathcal{V}$ into two sets $\mathbf{Z}$ and $\mathbf{W}$ with $\mathbf{X} \subseteq \mathbf{Z}$ and a setting $\mathbf{x}'$ and $\mathbf{w}$ of the variables in $\mathbf{X}$ and $\mathbf{W}$, respectively, such that $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}', \mathbf{W} \leftarrow \mathbf{w}]\neg\phi$.

AC2(b). For all subsets $\mathbf{Y}$ of $\mathbf{Z} - \mathbf{X}$, we have $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}, \mathbf{W} \leftarrow \mathbf{w}, \mathbf{Y} \leftarrow \mathbf{y}^*]\phi$.

We call $\mathbf{W} = \mathbf{w}$ a *witness* of $\mathbf{X} = \mathbf{x}$ causing $Y = y$.

Note that one choice of $\mathbf{Y}$ for which the condition in AC2(b) is required to hold, is $\mathbf{Y} = \emptyset$. For that choice, AC2 states that the effect counterfactually depends on the cause when holding fixed the witness $\mathbf{W} = \mathbf{w}$: $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}, \mathbf{W} \leftarrow \mathbf{w}]\phi$ and $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}', \mathbf{W} \leftarrow \mathbf{w}]\neg\phi$. Therefore AC2(a) can easily be interpeted as expressing a – contrastive – necessity condition: there exist contrast values $\mathbf{x}'$ such that if those values were to obtain, then AC2(b) no longer holds.

The problem lies with interpreting AC2(b) as expressing causal sufficiency. The main obstacle lies in the absence of the requirement that $\mathbf{w} = \mathbf{w}^*$, i.e., it is not required that the supposedly sufficient set of events $(\mathbf{X} = \mathbf{x}, \mathbf{W} = \mathbf{w})$ *actually took place*. Therefore we cannot simply view $(\mathbf{X} = \mathbf{x}, \mathbf{W} = \mathbf{w})$ itself as the causally sufficient set we are looking for. Although it cannot be excluded that the conditions imposed by invoking $\mathbf{Z}$ (and $\mathbf{Y}$) somehow ensure the existence of some other set that *can* be interpreted as a causally sufficient set, it is far from obvious that this is the case. This is confirmed by the fact that Halpern & Pearl do not even offer an attempt at giving an interpretation of AC2(b) as expressing causal sufficiency.

Matters get worse when we turn our attention to **Updated HP** [10]:

Definition 5 [**Updated HP**]

AC2(a). Identical to the previous one.

AC2(b). For all subsets $\mathbf{V}$ of $\mathbf{W}$ and subsets $\mathbf{Y}$ of $\mathbf{Z} - \mathbf{X}$, we have $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}, \mathbf{V} \leftarrow \mathbf{v}, \mathbf{Y} \leftarrow \mathbf{y}^*]\phi$ (where $\mathbf{v}$ is the restriction of $\mathbf{w}$ to $\mathbf{V}$).

We see that AC2(b) has become even more complicated, and yet no argument is given as to how this condition formalizes causal sufficiency, despite Halpern explicitly claiming that this is what it aims to do.[5] Instead, the updated version is justified on the basis of examples for which the previous version gave counterintuitive answers.

[5] Concretely, when discussing sufficient causality we find the following [8, p. 53]:

> The key intuition behind the definition of sufficient causality is that not only does $\mathbf{X} = \mathbf{x}$ suffice to bring about ϕ in the actual context (which is the intuition that AC2(b) [from **Original HP**] and AC2(b) [from **Updated HP**] are trying to capture)...

As a sidenote, Halpern and Pearl [10] also define *strong causation* by demanding that the following condition holds in addition to the other two:

AC2(c). For all $\mathbf{w} \in \mathcal{R}(\mathbf{W})$ we have that $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}, \mathbf{W} \leftarrow \mathbf{w}]\phi$.

This definition has received almost no attention in the literature, because according to Halpern & Pearl it is too strong.[6] As we shall see, this is unfortunate, because AC2(c) does adequately capture a variant of causal sufficiency.

Finally we have **Modified HP**, which is far simpler than the previous two [7].

Definition 6 [**Modified HP**]

AC2. There is a set $\mathbf{W}$ of variables in $\mathcal{V} - \mathbf{X}$, and a setting $\mathbf{x}'$ of the variables in $\mathbf{X}$ such that $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}', \mathbf{W} \leftarrow \mathbf{w}^*]\neg\phi$.

The crucial difference here is that **Modified HP** *does* require the witness to consist solely of events which actually took place, i.e., $\mathbf{w} = \mathbf{w}^*$. It is straightforward to show that simply adding this requirement ensures that both versions of AC2(b) are satisfied automatically, and therefore an explicit sufficiency condition is not required. Halpern considers this definition to be an improvemement over the other two, and I agree with him. However, Halpern arrives at this conclusion based on the many examples in which it better agrees with intuition. As will become clear, another – and arguably more compelling – justification is to be found in the fact that it is the only definition of the three which has a natural interpretation as formalizing the NESS intuition with which we started. To get there, we need to step away from the HP definitions and start afresh.

4 Causal Sufficiency

4.1 Some Technical Preliminaries

1: Halpern [8] suggests treating "part of a cause" (i.e., any $X = x$ that appears in $\mathbf{X} = \mathbf{x}$) as synonymous with "cause" when talking about **Modified HP**. I will follow this suggestion throughout whenever discussing the judgment of **Modified HP** in particular examples, unless stated otherwise. In stating theorems, however, the two are kept apart.

2: The HP definitions allow the effect to be any propositional formula ϕ, whereas the other definitions of causation will require effects to be of the form $Y = y$. A

[6] In retrospect, there is little basis for this judgment. They only discuss two examples in which strong causation diverges from **Updated HP**. In the first of those (Ex. 3.2), it fails to call the lighting of each of two matches ($ML_1 = 1$ and $ML_2 = 1$) to be causes of a forest fire, whereas **Updated HP** does not. However, their conjunction ($ML_1 = 1, ML_2 = 1$) *is* a strong cause, and thus each of them is part of a strong cause. As we will see, Halpern later suggests treating "part of a cause" as being synonymous to "cause", so the point would be moot. In the second example (Ex. 5.5), discussed as Example 4 later on, $S = 1$ is not a strong cause although it is a cause according to **Updated HP**. This is an example of *trumping causation*, for which the majority opinion is that $S = 1$ is indeed not a cause. Moreover, Halpern's later definition **Modified HP** also does not consider it a cause.

thorough discussion of complex effects is beyond the scope of this paper. I here limit myself to two observations.

- Although the definitions of causation here developed can be generalized to allow for conjunctive effects (i.e., effects of the form $\mathbf{Y} = \mathbf{y}$), it is not at all clear that we should want to do so. The reason is that we can easily include variables into the effect that have nothing whatsoever to do with the causes. Say we have a variable Y with equation $Y = U$, where U is an exogenous variable, and we are considering a context where $U = 1$. Then for any cause-effect pair $\mathbf{X} = \mathbf{x}$ and ϕ, we automatically get that $\mathbf{X} = \mathbf{x}$ also causes $\phi \wedge Y = 1$, which is not a sensible result. Therefore we choose to simply exclude conjunctive effects.
- In the few examples in the literature where the HP definitions actually consider an effect ϕ that is not of the form $Y = y$, ϕ takes on the form $Y = y_1 \vee Y = y_2, \ldots, \vee Y = y_n$ for some n. The definitions here developed can easily be generalized to also allow for such effects. For reasons of simplicitly I choose not to do so in general and limit the discussion of this generalization to one example for which it is required.

3: The definitions of sufficiency below (and the definitions of actual causation that follow in their wake) could be extended to also allow for exogenous variables as members of a sufficient set, so that exogenous and endogenous variables are treated alike. Since our goal is to make comparisons with the HP definitions, those would also have to be extended. Concretely, the HP definitions restrict causes to being endogenous variables, and they do not allow exogenous variables to be parts of a "witness" (the set $\mathbf{W}$ above). For example, if we have $Y = X \vee U$ where $U \in \mathcal{U}$ and we consider a context where $U = 1$ and $X = 1$, the HP definitions are unable to identify $X = 1$ as a cause because they disallow considering what happens when $U = 0$. The simplest way to sidestep this issue is to restrict ourselves to models where exogenous variables only appear in equations of the form $V = U$. In that manner, all influence of the exogenous variables can be overriden by interventions, reducing their role to simply providing us with the actual values of all variables. For any model which does not conform to this restriction, we can easily construct a very similar model that does: simply replace any exogenous variable U which appears in some equation that is not of this form with a new endogenous variable V_U, and add the equation $V_U = U$. For the previous example this results in the model with equations $Y = X \vee V_U$, $V_U = U$. (Note that now the HP definitions do consider $X = 1$ to be a cause of $Y = 1$.).

4.2 Six Variants of Sufficiency

Throughout the rest of the paper, we take $\mathbf{X}$ and $\mathbf{Y}$ to be non-identical subsets of the endogenous variables $\mathcal{V}$ that appear in a causal model M.[7]

[7]We take them to be non-identical to exclude calling a setting $\mathbf{X} = \mathbf{x}$ causally sufficient for itself, and a fortiori to exclude calling it a cause of itself. A reviewer pointed out to me that Halpern and Pearl [10] do not rule out self-causation, although they did consider doing so.

Informally, to say that some setting $\mathbf{X} = \mathbf{x}$ is sufficient for another setting $\mathbf{Y} = \mathbf{y}$, is to say that the latter follows from the former.[8] To formalize this requires making explicit what it means for one setting to "follow" from another. In the context of *causal* sufficiency, an obvious minimal demand is that this meaning captures the causal directionality. In the framework of causal models this comes down to treating $\mathbf{X} = \mathbf{x}$ as an intervention and $\mathbf{Y} = \mathbf{y}$ as a consequence of that intervention: if we set $\mathbf{X}$ to the values $\mathbf{x}$, then $\mathbf{Y}$ takes on the values $\mathbf{y}$. At least this much is clear.

Yet by saying this, we have said nothing at all about the other endogenous variables and their values, nor about the contexts in which we are evaluating the intervention. The difficulty lies in deciding what conditions we choose to impose on the other variables, both endogenous and exogenous. I consider six possible ways in which this decision can be made that are fairly natural, but this is by no means an exhaustive list.

We start with the strongest conditions possible: *in all contexts*, if we set $\mathbf{X}$ to the values $\mathbf{x}$, then $\mathbf{Y}$ takes on the values $\mathbf{y}$, *independent of the values of all other variables*.[9]

Definition 7 We say that $\mathbf{X} = \mathbf{x}$ *is directly sufficient for* $\mathbf{Y} = \mathbf{y}$ *in* M if for all $\mathbf{c} \in \mathcal{R}(\mathcal{V} - (\mathbf{X} \cup \mathbf{Y}))$ and all $\mathbf{u} \in \mathcal{R}(\mathcal{U})$ we have that $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}, \mathbf{C} \leftarrow \mathbf{c}]\mathbf{Y} = \mathbf{y}$.

The strength of this definition is also its weakness: by putting such strong demands on the sufficient set, many interesting sets are excluded. This restrictiveness becomes apparent later on when we add a necessity condition (Proposition 5): only parents can ever be part of a minimal directly sufficient set. A trivial example illustrates this point. Say the equation for Y is $Y = A$, the equation for A is $A = X$, and we are looking at a context in which $X = 1$.[10] Then $X = 1$ is not directly sufficient for $Y = 1$, because intervening on A overrides any influence of X on Y. Still, there is clearly a sense in which $X = 1$ *is* causally sufficient for $Y = 1$. In particular, $X = 1$ is directly sufficient for $(A = 1, Y = 1)$.

Generalizing this intuition provides us with the second form of sufficiency: there is some setting $\mathbf{N} = \mathbf{n}$ that includes $\mathbf{Y} = \mathbf{y}$, so that in all contexts, if we set $\mathbf{X}$ to the values $\mathbf{x}$, then $\mathbf{N}$ takes on the values $\mathbf{n}$, independent of the values of all other variables. This can be formulated more succinctly as: $\mathbf{X} = \mathbf{x}$ is directly sufficient for some set to which $\mathbf{Y} = \mathbf{y}$ belongs.

Definition 8 We say that $\mathbf{X} = \mathbf{x}$ *is strongly sufficient for* $\mathbf{Y} = \mathbf{y}$ *in* M if there exists a $\mathbf{N} = \mathbf{n}$ so that $\mathbf{Y} \subseteq \mathbf{N}$, $\mathbf{y}$ is the restriction of $\mathbf{n}$ to $\mathbf{Y}$, and $\mathbf{X} = \mathbf{x}$ is directly sufficient for $\mathbf{N} = \mathbf{n}$.

[8] Note that in this paper we are interested in the causal sufficiency of *settings of variables* for other settings of variables. This is quite distinct from how the term "causal sufficiency" is sometimes used in the causal modelling literature, namely as a property of a *set of variables* in a causal graph.

[9] Weslake [20] also offers this definition of causal sufficiency to develop a definition of actual causation. He mistakenly claims that Halpern & Pearl call this condition strong causation. As we have seen, strong causation does not require $\mathbf{C}$ to contain *all* other variables.

[10] In all examples the variables are binary unless indicated otherwise. A binary variable is a variable that has range $\{0, 1\}$.

Observe that another intuitive way of viewing $X = 1$ as being causally sufficient for $Y = 1$ in the simple example we just discussed, is to note that $X = 1$ is directly sufficient for $A = 1$ and $A = 1$ is directly sufficient for $Y = 1$. This intuition can also be generalized to define a form of sufficiency. Concretely, we can define *strong sufficiency along a network* as the transitive closure of direct sufficiency.[11]

Definition 9 We say that $\mathbf{X} = \mathbf{x}$ *is strongly sufficient for* $\mathbf{Y} = \mathbf{y}$ *in* M *along a network* $\mathbf{N}$ if there are (possibly overlapping) sets $\mathbf{N_i}$ such that $\mathbf{N} = \mathbf{Y} \cup_{i \in \{1,...,k\}} \mathbf{N_i}$ and there exist values $\mathbf{n_i} \in \mathcal{R}(\mathbf{N_i})$ for each i such that $\mathbf{X} = \mathbf{x}$ is directly sufficient for $\mathbf{N_1} = \mathbf{n_1}$, $\mathbf{N_1} = \mathbf{n_1}$ is directly sufficient for $\mathbf{N_2} = \mathbf{n_2}$, ..., and $\mathbf{N_k} = \mathbf{n_k}$ is directly sufficient for $\mathbf{Y} = \mathbf{y}$.

The following result shows that both forms of strong sufficiency are merely different ways of expressing the same notion of sufficiency (and hence the term is appropriately chosen). Taking in mind the earlier observation (to appear later as Proposition 5) that direct sufficiency combined with necessity is a relation between parents and children, we can safely think of a network as consisting of variables that lie on some path between $\mathbf{X}$ and $\mathbf{Y}$. Doing so will make it easier to apply the definitions of causation to examples.

Proposition 1 $\mathbf{X} = \mathbf{x}$ *is strongly sufficient for* $\mathbf{Y} = \mathbf{y}$ *in* M *along a network* $\mathbf{N}$ *iff* $\mathbf{X} = \mathbf{x}$ *is strongly sufficient for* $\mathbf{Y} = \mathbf{y}$ *in* M.

(Proofs of all Theorems are to be found in the Appendices A, B, C and D.)

Another obvious way to weaken the conditions on the values of the endogenous variables compared to direct sufficiency is to only consider the setting in which we leave the other variables alone, giving: *in all contexts*, if we set $\mathbf{X}$ to the values $\mathbf{x}$ *and do not intervene on any other variable*, then $\mathbf{Y}$ takes on the values $\mathbf{y}$.[12]

Definition 10 We say that $\mathbf{X} = \mathbf{x}$ *is weakly sufficient for* $\mathbf{Y} = \mathbf{y}$ *in* M if for all $\mathbf{u} \in \mathcal{R}(\mathcal{U})$ we have that $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}]\mathbf{Y} = \mathbf{y}$.

The following straightforward result shows the relative strengths of the above three notions of sufficiency.

Proposition 2 *If* $\mathbf{X} = \mathbf{x}$ *is directly sufficient for* $\mathbf{Y} = \mathbf{y}$ *then* $\mathbf{X} = \mathbf{x}$ *is strongly sufficient for* $\mathbf{Y} = \mathbf{y}$*, and if* $\mathbf{X} = \mathbf{x}$ *is strongly sufficient for* $\mathbf{Y} = \mathbf{y}$ *then* $\mathbf{X} = \mathbf{x}$ *is weakly sufficient for* $\mathbf{Y} = \mathbf{y}$.

[11] As with the definition of direct sufficiency, this one also appears in Weslake [20]'s construction of actual causation, with the added requirement that $\mathbf{N}$ is minimal. This demand becomes redundant once we add our necessity condition. The other conditions Weslake invokes are quite complicated and do not have a counterpart in our story, which is why his definition also fails at the first strategy.

[12] This definition appears as just one condition in Halpern [8]'s definition of *sufficient causality*. One of the other conditions is in fact actual causation.

So far we have considered three definitions that differ only with regards to the conditions they impose on the values of the endogenous variables: they all agreed on requiring their respective conditions to hold in all contexts. Yet questions of actual causation are posed relative to an actual context $\mathbf{u}$, and thus it is only natural that we should consider doing the same for questions of causal sufficiency. This adds three more definitions of sufficiency, which are simply the result of replacing the universal quantifier over contexts with a particular context that is assumed to be given.

Definition 11 We say that $\mathbf{X} = \mathbf{x}$ *is actually directly sufficient for* $\mathbf{Y} = \mathbf{y}$ *in* $(M, \mathbf{u})$ if for all $\mathbf{c} \in \mathcal{R}(\mathcal{V} - (\mathbf{X} \cup \mathbf{Y}))$ we have that $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}, \mathbf{C} \leftarrow \mathbf{c}]\mathbf{Y} = \mathbf{y}$.

Definition 12 We say that $\mathbf{X} = \mathbf{x}$ *is actually strongly sufficient for* $\mathbf{Y} = \mathbf{y}$ *in* $(M, \mathbf{u})$ if there exist $\mathbf{N} = \mathbf{n}$ so that $\mathbf{Y} \subseteq \mathbf{N}$, $\mathbf{y}$ is the restriction of $\mathbf{n}$ to $\mathbf{Y}$, and $\mathbf{X} = \mathbf{x}$ is actually directly sufficient for $\mathbf{N} = \mathbf{n}$.

Definition 13 We say that $\mathbf{X} = \mathbf{x}$ *is actually weakly sufficient for* $\mathbf{Y} = \mathbf{y}$ *in* $(M, \mathbf{u})$ if $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}]\mathbf{Y} = \mathbf{y}$.

Obviously the counterpart of Proposition 2 holds as well for these notions of actual sufficiency.

4.3 General Form of Causal Sufficiency

We can formalize and generalize the intuitions behind the definitions in the preceding section by showing that all six definitions of sufficiency can be interpreted as simply putting different constraints on the parameters that occur in the following general definition of sufficiency. (We only explicitly discuss the three definitions of "non-actual" sufficiency, but the same analysis trivially applies to the three definitions of actual sufficiency.)

Definition 14 [**General Definition of Sufficiency**] We say that $\mathbf{X} = \mathbf{x}$ is *sufficient* for $\mathbf{Y} = \mathbf{y}$ in M if there exist sets $\mathbf{C} \subseteq \mathcal{V} - (\mathbf{X} \cup \mathbf{Y})$, $\mathbf{N} \subseteq \mathcal{V} - (\mathbf{X} \cup \mathbf{C})$ with $\mathbf{Y} \subseteq \mathbf{N}$, and a setting $\mathbf{n} \in \mathcal{R}(\mathbf{N})$ where the restriction of $\mathbf{n}$ to $\mathbf{Y}$ is $\mathbf{y}$, such that for all $\mathbf{c} \in \mathcal{R}(\mathbf{C})$ and for all $\mathbf{u} \in \mathcal{R}(\mathcal{U})$ we have that $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}, \mathbf{C} \leftarrow \mathbf{c}]\mathbf{N} = \mathbf{n}$.

We say that $\mathbf{X} = \mathbf{x}$ is sufficient for $\mathbf{Y} = \mathbf{y}$ in M along $\mathbf{N}$ independent of $\mathbf{C}$.

This definition is more complicated than Definitions 7, 8, and 10. Its use lies in the fact that it allows us to see exactly how the three definitions relate to each other, and how one can construct other definitions of sufficiency, by invoking the following trivial result.

Proposition 3 Definitions 7, 8, *and* 10, *are equivalent to* Definition 14 *when making respectively the following choices for* $\mathbf{N}$ *and* $\mathbf{C}$*:*

Weak Sufficiency. *Choose both* $\mathbf{C}$ *and* $\mathbf{N}$ *to be minimal, i.e.,* $\mathbf{C} = \emptyset$, $\mathbf{N} = \mathbf{Y}$.
Strong Sufficiency. *Choose* $\mathbf{N}$ *to be maximal given* $\mathbf{C}$*, i.e.,* $\mathbf{N} = \mathcal{V} - (\mathbf{X} \cup \mathbf{C})$.

Direct Sufficiency. *Choose* $\mathbf{C}$ *to be maximal, i.e.,* $\mathbf{C} = \mathcal{V} - (\mathbf{X} \cup \mathbf{Y})$ *and thus* $\mathbf{N} = \mathbf{Y}$.

Proposition 3 could inspire even more variants of sufficiency. In fact, we have already come across the most obvious one: AC2(c). It is easy to see that it consists of choosing $\mathbf{N}$ to be minimal given $\mathbf{C}$, i.e., $\mathbf{N} = \mathbf{Y}$, meaning it sits in between Weak and Strong Sufficiency. The condition also appears as a sufficiency condition in Pearl's notion of *sustenance*, which is the first step he takes towards formalizing the NESS intuition [17]. Unfortunately it is also the last step, because the subsequent notions he introduces are far more complicated and bear no resemblance to NESS. The added complexity is introduced precisely because taken by itself sustenance fails to provide a sensible definition of causation, which is why I leave the exploration of this and other possible variants of sufficiency for another occasion.

5 Defining Causation Using Sufficiency

We are finally ready to take up the main challenge: defining actual causation as the formal expression of the NESS intuition. In order to do so, several questions need to be answered:

- Should we use actual sufficiency or not?
- Which of the three definitions of (actual) causal sufficiency should we use?
- Does necessity mean that there exist contrast values of $\mathbf{X}$ so that the set would not be sufficient if those values obtained, or does it mean that the set is no longer sufficient when we remove the subset $\mathbf{X}$?

I have introduced six definitions of causal sufficiency in the previous section. For each definition, we can define causation using either of the two interpretations of necessity, giving twelve definitions of actual causation altogether. However, I will show that several of these are equivalent to each other, and one will be impossible to satisfy, leaving us with six definitions in the end. One of those will be **Modified HP**.

5.1 A Family of Definitions

As with the HP definitions, Definition 3 gives the general form of all definitions, except that ϕ is restricted to $Y = y$. (This restriction is assumed whenever comparisons are made with the HP definitions.) As before, the only difference lies with the content of AC2. Using the first interpretation of necessity, which we shall call *contrastive necessity*, the general form of AC2 is as follows:

Definition 15 [**General Definition of Causation**] There exist sets $\mathbf{W}$, $\mathbf{N}$ such that

AC2(a^{c}). There exist values $\mathbf{x}'$ such that for all $\mathbf{S} \subseteq \mathbf{N}$, $(\mathbf{X} = \mathbf{x}', \mathbf{W} = \mathbf{w}^*)$ is not sufficient for $Y = y$ along $\mathbf{S}$.

AC2(b). $(\mathbf{X} = \mathbf{x}, \mathbf{W} = \mathbf{w}^*)$ is sufficient for $Y = y$ along $\mathbf{N}$.

We call $\mathbf{W}$ a *witness* of $\mathbf{X} = \mathbf{x}$ causing $Y = y$.

By replacing sufficiency in the **General Definition of Causation** with any of the six definitions of sufficiency from Section 4, we obtain six specific definitions of actual causation.[13] AC2(b) simply expresses causal sufficiency, whatever form it may take. AC2(a^c) offers a somewhat nuanced expression of necessity because it also focusses on subsets of $\mathbf{N}$. (Note that this nuance matters only for Strong Sufficiency, since for Weak and Direct Sufficiency $\mathbf{N} = \{Y\}$ anyway.) The reason is that our interest lies with the sufficiency for $Y = y$, and the network $\mathbf{N}$ is merely a means to that end. If $\mathbf{X} = \mathbf{x}'$ accomplishes the same end using less means, then $\mathbf{X} = \mathbf{x}$ was not necessary for achieving it.

Under the second interpretation of necessity, which we shall call *minimal necessity*, AC2(a^c) is replaced with:

AC2(a^m). For all $\mathbf{S} \subseteq \mathbf{N}$, $\mathbf{W} = \mathbf{w}^*$ is not sufficient for $Y = y$ along $\mathbf{S}$.

Both interpretations of necessity are prima facie plausible. The contrastive interpretation is explicitly counterfactual in nature, whereas the minimal interpretation is more neutral. Our analysis will settle which one of them is to be preferred.

Filling in each of the six definitions of causal sufficiency into both versions of the **General Definition of Causation** gives twelve specific definitions of actual causation. I refer to each of these as **Def** x for $x \in \{1, \ldots, 12\}$ along the following convention:

- **Def 1** Contrastive actual weak sufficiency
- **Def 2** Contrastive actual strong sufficiency
- **Def 3** Contrastive actual direct sufficiency
- **Def 4** Contrastive weak sufficiency
- **Def 5** Contrastive strong sufficiency
- **Def 6** Contrastive direct sufficiency
- **Def 7** Minimal actual weak sufficiency
- **Def 8** Minimal actual strong sufficiency
- **Def 9** Minimal actual direct sufficiency
- **Def 10** Minimal weak sufficiency
- **Def 11** Minimal strong sufficiency
- **Def 12** Minimal direct sufficiency

So to be clear, each **Def** x is constructed by taking the respective definition of sufficiency (i.e., Definitions 7, 8, 10, 11, 12, or 13), filling that into the **General Definition of Causation** where AC2(a) takes on AC2(a^c) or AC2(a^m) depending on whether $x < 7$ or not, and finally, filling those conditions AC2 into Definition 3. I illustrate the result of this construction for **Def 2**.

Definition 16 [**Def 2**] $\mathbf{X} = \mathbf{x}$ is an *actual cause* of $Y = y$ according to **Def 2** in $(M, \mathbf{u})$ if the following three conditions hold:

[13] Definition 15 can be made even more general by also incorporating $\mathbf{C}$ from Definition 14. Since we are only considering notions of sufficiency for which $\mathbf{C}$ is determined entirely by the other sets, there is no need to do so for our purposes. But it is important to keep this additional generality in mind if one wants to use alternative definitions of sufficiency.

AC1. $(M, \mathbf{u}) \models (\mathbf{X} = \mathbf{x}) \wedge Y = y$.

AC2(a^c). There exist sets $\mathbf{W}$, $\mathbf{N}$ with $Y \in \mathbf{N}$, and values $\mathbf{x}'$, such that for all $\mathbf{S} \subseteq \mathbf{N}$ with $Y \in \mathbf{S}$, and for all $\mathbf{s} \in \mathcal{R}(\mathbf{S})$ such that $y \in \mathbf{s}$, there exists a $\mathbf{t} \in \mathcal{R}(\mathcal{V} - (\mathbf{X} \cup \mathbf{W} \cup \mathbf{S}))$ so that $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}', \mathbf{W} \leftarrow \mathbf{w}^*, \mathbf{T} \leftarrow \mathbf{t}]\mathbf{S} \neq \mathbf{s}$.

AC2(b). For all $\mathbf{c} \in \mathcal{R}(\mathcal{V} - (\mathbf{X} \cup \mathbf{W} \cup \mathbf{N}))$ we have that $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}, \mathbf{W} \leftarrow \mathbf{w}^*, \mathbf{C} \leftarrow \mathbf{c}]\mathbf{N} = \mathbf{n}^*$.

AC3. $\mathbf{X}$ is minimal.

Admittedly, **Def 2** looks even more complicated than **Updated HP**. Further on I provide some results that allow us in many cases to use simpler definitions as stand-ins for **Def 2**. More importantly, although the notation of Definition 16 is complicated, its meaning can be spelled out intuitively by stating that $\mathbf{X} = \mathbf{x}$ causes $Y = y$ iff $\mathbf{X} = \mathbf{x}$ is a Minimal Contrastively Necessary Subset of a Strongly Sufficient Set for $Y = y$ (or MCNS^4).[14]

5.2 Analysis

Let us now turn to investigating the relations between these definitions. (Knowing these relations before getting into the discussion of examples makes life a lot easier.) A first remark is that **Def 7** is impossible to satisfy, as it requires that both $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}^*, \mathbf{W} \leftarrow \mathbf{w}^*]Y = y$ and $(M, \mathbf{u}) \not\models [\mathbf{W} \leftarrow \mathbf{w}^*]Y = y$ hold, implying that $(M, \mathbf{u}) \models Y = y \wedge Y \neq y$.

A second remark is that **Def 3** is equivalent to a condition that appears in Pearl's first definition of actual causation [15].[15]

Ignoring **Def 7**, we are still left with eleven candidate definitions of actual causation (fourteen candidates if we count the three HP definitions), whereas we would like to settle on just one. The rest of the paper is concerned with selecting the best definition out of the lot. As a first step, we can reduce the number of definitions by six.

Theorem 1 *The following are all equivalences among the twelve definitions and the three HP definitions:*

- **Modified HP** *iff* **Def 1**
- **Def 2** *iff* **Def 5**
- **Def 8** *iff* **Def 11**
- **Def 3** *iff* **Def 6** *iff* **Def 9** *iff* **Def 12**

Theorem 1 offers our first interesting result: it shows that **Modified HP** succeeds in formalizing the NESS intuition, whereas the other two HP definitions do not. From now on I will ignore the definitions appearing on the right-hand side in Theorem 1.

[14] Strictly speaking it should say "Actually Strongly Sufficient", but that makes for a less elegant acronym. I am cheating a bit by anticipating Theorem 1.

[15] It re-appears in his second definition of actual causation in the notion of a *causal beam*, but without the necessity condition [17, p. 318]. To see the equivalence, one needs to invoke Proposition 5.

The following is a helpful result for applying some of the definitions going forward. (As is well known, the same result holds for **Original HP** [8].)

Proposition 4 *If* $\mathbf{X} = \mathbf{x}$ *causes* $Y = y$ *in* $(M, \mathbf{u})$ *according to a definition that uses minimal necessity, then* $\mathbf{X}$ *is a singleton.*

The following result offers important insights into the relations between the remaining definitions.

Theorem 2 *The only implications – involving either causes or parts of causes – between the remaining five definitions* (**Def 2**, **Def 3**, **Def 4**, **Def 8**, *and* **Def 10**) *and the three HP definitions are the following ones (and their immediate consequences, of course):*

- *If part of* **Modified HP** *then* **Updated HP**;[16]
- *If part of* **Updated HP** *then* **Original HP**;
- *If* **Def 3** *then* **Def 2**;
- *If part of* **Def 2** *then* **Def 8**;
- *If* **Def 3** *then* **Original HP**;
- *If* **Def 10** *then* **Def 4**.

6 Excluding Def 3 and Def 10

Two definitions can be excluded quickly. The following result shows why **Def 3** is not a sensible candidate as a general definition of causation, since causation is obviously not restricted to parent-children pairs.

Proposition 5 *If* $\mathbf{X} = \mathbf{x}$ *causes* $Y = y$ *in* $(M, \mathbf{u})$ *according to* ***Def 3****, then* $\mathbf{X}$ *is a singleton, and* X *is a parent of* Y.

Although we can dismiss **Def 3** as a general definition of causation, it is still a useful stand-in for – the arguably more complicated – **Def 2** and **Def 8** in case X is a parent of Y and X is not an ancestor of Y along any path that is longer than a single edge (which in fact covers a surprisingly large number of cases discussed in the literature). In such cases we say that X is *only* a parent of Y.

Proposition 6 *If* X *is only a parent of* Y*, then* **Def 2**, **Def 3**, *and* **Def 8** *are all equivalent for causes* $X = x$.

A cornerstone of the counterfactual approach to causation is that counterfactual dependence is sufficient for causation. More formally, there is widespread consensus that causation should satisfy the following principle:[17]

[16] This is shorthand for: If $X = x$ is part of a cause of $Y = y$ according to the **Modified HP** definition then it is a cause of $Y = y$ according to the **Updated HP** definition.

[17] As does Halpern, I here restrict myself to counterfactual dependence on a single conjunct [8].

Principle 1 (**Dependence**) *Say* $(M, \mathbf{u}) \models X = x \wedge Y = y$. *If there exists a value* x' *such that* $(M, \mathbf{u}) \models [X \leftarrow x']Y \neq y$ *then* $X = x$ *causes* $Y = y$ *in* $(M, \mathbf{u})$.

Accepting this principle means that **Def 10** is excluded as well.

Proposition 7 *Out of all definitions we have considered,* **Def 10** *and* **Def 3** *are the only ones which do not satisfy* **Dependence**.

That leaves us with **Def 2**, **Def 4**, and **Def 8** as possible alternatives to the HP definitions.

7 Def 2, Def 4, and Def 8, vs the HP Definitions

We have shown that all twelve definitions we developed (including **Modified HP**) are instantiations of the **General Definition of Causation** (Def. 15), and thereby they improve upon **Original HP** and **Updated HP** as far as the first strategy goes. We now show that **Def 2** also improves upon all three HP definitions as far as the second strategy goes, whereas **Def 4** and **Def 8** do not. In order to remain as neutral as possible, we go over Halpern & Pearl's own examples, compare the verdicts of our definitions to theirs, and stick as close as possible to their intuitions.

7.1 Comparison to Updated HP

The **Updated HP** definition is by far the most well-known. It was developed as an improvement of **Original HP**, which sometimes gives unreasonable answers. Halpern and Pearl [10] offer many examples to illustrate how it works and how it successfully deals with paradigm cases of causation.

Their first example is one of those few cases – recall the beginning of Section 4 – in which the effect is of the form $Y = y_1 \vee Y = y_2$, and therefore allows us to illustrate how we can generalize the **General Definition of Causation** to such effects. It is also an example for which **Def 8** gives the wrong answer, but the subsequent example is far simpler and more convincing in this respect.

Example 1 "Suppose that there was a heavy rain in April and electrical storms in the following two months; and in June the lightning took hold. If it hadn't been for the heavy rain in April, the forest would have caught fire in May." [10, p. 15] I agree with Halpern and Pearl's judgment that it would be very counterintuitive to say that the April rain caused the forest fire, since all it did was delay the fire. As they indicate, it is nevertheless perfectly sensible to say that the April rain caused the forest fire *in June*, as opposed to May. In order to capture this distinction, we need to invoke a disjunctive effect.

Let F represent there being a fire or not, with three possible values: 0 (no fire), 1 (fire in May), or 2 (fire in June). ES is a four-valued variable that captures whether there are electric storms: $(0, 0)$ (no electric storms in either May or June), $(1, 0)$ (electric storms in May but not in June), $(0, 1)$ (storms in June but not May), and $(1, 1)$ (storms in both May and June). Lastly, AS is a binary variable expressing whether or not there was April rain.

The equation for F is then given by: $F = 2$ if $(AS = 1 \wedge ES = (1, 1)) \vee ES = (0, 1)$, $F = 1$ if $AS = 0 \wedge (ES = (1, 1) \vee ES = (1, 0))$, and $F = 0$ otherwise. Given that $F = 2$ counterfactually depends on $AS = 1$, all definitions we are considering agree that $AS = 1$ causes $F = 2$. The question is whether $AS = 1$ also caused there to be a fire, i.e., whether it caused $F = 1 \vee F = 2$.

We can easily generalize sufficiency to such disjunctions: $\mathbf{X} = \mathbf{x}$ is sufficient for $Y = y \vee Y = y'$ iff $\mathbf{X} = \mathbf{x}$ is sufficient for $Y = y$ or $\mathbf{X} = \mathbf{x}$ is sufficient for $Y = y'$.[18] When integrated into our **General Definition of Causation**, this results in splitting up AC2(a) so that there is one instance for each disjunct. AC2(b) need not be split up, since it can only ever be satisfied for the actual value of Y.[19]

Let us apply this idea to our example. To satisfy AC2(b), we have to add ES to the witness: $(AS = 1, ES = (1, 1))$ is directly sufficient for $F = 2$ and $AS = 1$ is not. (We can focus on direct sufficiency because AS is only a parent of F. We cannot invoke Proposition 6 though, since that requires an effect $Y = y$.) We then see that one of the two conditions that now make up AC2(a) is not satisfied for **Def 2** and **Def 4**, because $(AS = 0, ES = (1, 1))$ is directly sufficient for $F = 1$. Therefore **Def 2** and **Def 4** agree with the HP definitions that the April rain did not cause the forest fire. But **Def 8** does not reach this verdict, because $ES = (1, 1)$ is not directly sufficient for either $F = 1$, nor is it for $F = 2$. This means AC2(a) is fullfilled for **Def 8**, which leads to a mistaken conclusion.

Although one counterexample need not disqualify a definition, the following example is indicative of a deeper problem with **Def 8**: whenever $X = x$ strongly suffices for $Y = y$, it is automatically a cause according to **Def 8**, since $\emptyset$ is never strongly sufficient for $Y = y$. The following example is but one of many paradigm cases in the literature for which this property leads to a counterintuitive verdict.[20] Therefore **Def 8** is also excluded as a definition of causation.

[18] A reviewer pointed out the following worry with this proposal. Imagine there are only two contexts, $\mathbf{u}$ and $\mathbf{u}'$, and we have that $\mathbf{X} = \mathbf{x}$ is sufficient for $Y = y$ in $(M, \mathbf{u})$ and $\mathbf{X} = \mathbf{x}$ is sufficient for $Y = y'$ in $(M, \mathbf{u}')$. If we then move to non-actual sufficiency, we have to quantify over all contexts, and thus we get that $\mathbf{X} = \mathbf{x}$ is not sufficient for $Y = y$ in M and nor is $\mathbf{X} = \mathbf{x}$ sufficient for $Y = y'$ in M. This means that under the current proposal it would not follow that $\mathbf{X} = \mathbf{x}$ is sufficient for $Y = y \vee Y = y'$ in M, which may seem counterintuitive to some. For the current purposes we can dismiss this worry as it is irrelevant for the following two reasons. First, Lemmas 1 and 2 together with the subsequent discussion in the proof of Theorem 1 show that as far as applying the **General Definition of Causation** is concerned, actual and non-actual sufficiency are equivalent for both direct and strong sufficiency. Therefore the imagined situation cannot arise for any candidate cause $\mathbf{X} = \mathbf{x}$ and effect $Y = y$. Second, as I will argue in favor of adopting a definition that uses strong sufficiency, I am content with setting aside the remaining worry one might have with regards to weak sufficiency.

[19] Note that this means generalizing to disjunctions across different variables – i.e., something like $Y = y \vee Z = z$ – is more complicated.

[20] McDermott [14] offers an almost identical example involving a dog biting a terrorist. Another famous case is that involving a boulder rolling towards a hiker [11]. All of these examples are counterexamples to the transitivity of causation. The failure of transitivity has become broadly accepted by now [2]. Despite what **Def 8**'s behavior in these examples might suggest, it is also not transitive. A simple counterexample consists of equations $Z = Y \vee W$, and $Y = X \wedge W$. If $X = W = 1$, **Def 8** considers $X = 1$ a cause of $Y = 1$, $Y = 1$ a cause of $Z = 1$, yet it does not consider $X = 1$ a cause of $Z = 1$.

Example 2 "The engineer is standing by a switch in the railroad tracks. A train approaches in the distance. She flips the switch, so that the train travels down the right-hand track, instead of the left. Since the tracks reconverge up ahead, the train arrives at its destination all the same...

Again, our causal model gets this right. Suppose we have three random variables:

- F for "flip", with values 0 (the engineer doesn't flip the switch) and 1 (she does);
- T for "track", with values 0 (the train goes on the left-hand track) and 1 (it goes on the right-hand track); and
- A for "arrival", with values 0 (the train does not arrive at the point of reconvergence) and 1 (it does).

" [10, p. 26]

First observe that as described, this causal model makes little sense: the equation for A is given by $A = T \vee \neg T$, which can be rewritten as $A = 1$. This can be fixed by extending the range of T with a value 2, representing the train not going down any track (because it breaks down, for example). Then the equations become $A = (T \neq 2)$ and $T = F$. The context is such that $F = 1$.

$F = 1$ is both weakly sufficient for $A = 1$ and strongly sufficient for $A = 1$ along $\{T\}$, but so is $F = 0$. Therefore **Def 2** and **Def 4** agree with **Updated HP** (and with intuition) that flipping the switch is not a cause of the train's arrival. **Def 8** fails to reach this verdict, because $\emptyset$ is not strongly sufficient for $A = 1$.

Def 4 suffers from an even bigger defect than **Def 8**: it fails to distinguish preempted causes from preempting causes. Since preemption cases are the bread and butter of the literature on actual causation, this means that **Def 4** is immediately disqualified. The following is a famous example of late preemption discussed by Halpern and Pearl [10] (and originally by Hall [5]).

Example 3 Suzy and Billy both throw a rock at a bottle. Suzy's rock gets there first, shattering the bottle. However Billy's throw was also accurate, and would have shattered the bottle had it not been preempted by Suzy's throw. Halpern & Pearl [10] use the following variables for this example, which capture the fact that Billy's throw was preempted by Suzy's rock hitting the bottle: BS for the bottle shattering, BH, SH for Billy's (resp. Suzy's) rock hitting the bottle, and two more variables (BT, ST) for either of them throwing their rock. The equations are then as follows: $BS = BH \vee SH$, $SH = ST$, $BH = BT \wedge \neg SH$. None of the definitions has any problem arriving at the obvious result that Suzy's throw ($ST = 1$) causes the bottle to shatter ($BS = 1$). However, **Def 4** is the only definition under consideration that mistakenly also judges Billy's throw to be a cause of the bottle's shattering: in all contexts $BT = 1$ is weakly sufficient for $BS = 1$, whereas $BT = 0$ is not weakly sufficient for $BS = 1$ in the context where $ST = 0$.

This leaves us with **Def 2** as the last potential alternative to the HP definitions. Going through the many remaining examples, there is only one in which **Def 2** disagrees with **Updated HP**. I leave it to the reader to verify this claim, and restrict the discussion to that single example.

Example 4 Major (M) and sergeant (S) stand before corporal, and both shout 'Charge!' ($M = 1$, $S = 1$). The corporal charges ($C = 1$). Orders from higher-ranking soldiers trump those of lower rank, so if the major had shouted 'Halt' ($M = 0$) the corporal would not have charged. If the major remains quiet ($M = -1$), the corporal listens to the sergeant.[21] The equation for C is thus: $C = M$ if $M \neq -1$ and $C = S$ otherwise. The majority intuition is that the sergeant did not cause the corporal to charge, because his order was trumped by that of the major.[22]

Def 2 agrees, as it does not consider $S = 1$ a cause of $C = 1$. The reason is that $M = 1$ is directly sufficient by itself, and yet $S = 1$ needs $M = 1$ as a witness to form a sufficient set. $S = 1$ is a cause of $C = 1$ according to both **Original HP** and **Updated HP**. Halpern & Pearl do not consider this to be problematic, but they do go through the trouble of showing how **Original HP** and **Updated HP** change their verdict if one adds extra variables to the model. Moreover, **Modified HP** also agrees with **Def 2** here. Given Halpern's later preference for **Modified HP**, it is fair to say that **Def 2** does at least as good as **Updated HP** on this example.

7.2 Comparison to Modified HP

Dissatisfied with **Updated HP** due to the many counterexamples that were presented in the literature, Halpern [7] develops **Modified HP**. First of all, despite Theorem 2, there do exist interesting connections between the three definitions we have considered and **Modified HP**.

Proposition 8 *If* **Modified HP** *with* **X** *a singleton, then* **Def 2**, **Def 4**, *and* **Def 8**.

Halpern [7] goes over several counterexamples to **Updated HP** and shows that **Modified HP** offers sensible verdicts. Taking into account Halpern's suggestion that "part of cause" is synonymous with "cause" for **Modified HP**, there are in fact only three examples in which **Modified HP** disagrees with **Updated HP** (Examples 3.5, 3.8, and 3.11).[23] In all three of those cases, **Def 2** sides with **Modified HP**.

There is only one example in which **Def 2** disagrees with **Modified HP**.[24] Crucially, it is an example for which Halpern agrees that **Modified HP** reaches the wrong verdict.

Example 5 A ranch has five individuals: $a_1, \ldots, a_5$. They have to vote on two possible outcomes: staying at the campfire ($O = 0$) or going on a round-up ($O = 1$). Let

[21] This formulation is due to Weslake [20], but the example was first discussed by Schaffer [19] (who attributes it to van Fraassen).

[22] See Weslake [20] for a discussion.

[23] When discussing Example 3.8 again in Halpern [8], he mistakenly claims that **Modified HP** agrees with **Updated HP** when treating parts of causes as causes. In response, Halpern has suggested a small variation on the example in which **Modified HP** indeed does agree with **Updated HP** (personal communication). For that variation, **Def 2** also agrees with the HP definitions.

[24] Halpern [8] discusses far more cases, but none of them reveal any further disagreements between these definitions.

A_i be the random variable denoting a_i's vote, so $A_i = j$ if a_i votes for outcome j.[25] There is a complicated rule for deciding on the outcome. If a_1 and a_2 agree (i.e., if $A_1 = A_2$), then that is the outcome. If $a_2, \ldots, a_5$ agree, and a_1 votes differently, then the outcome is given by a_1's vote (i.e., $O = A_1$). Otherwise, majority rules. In the actual situation, $A_1 = A_2 = 1$ and $A_3 = A_4 = A_5 = 0$, so by the first mechanism, $O = 1$.[26]

Halpern states, and I agree, that intuitively one should expect only $A_1 = 1$ and $A_2 = 1$ to be causes of $O = 1$. After all, $a_3, \ldots, a_5$ voted *against* $O = 1$. **Def 2** gives that result, whereas **Modified HP** considers every vote to be a cause. Halpern argues for adding more variables to the model in order to get the right outcome, but it speaks in favor of **Def 2** that it is able to give the right answer with just these variables.

We conclude that judged by the second strategy and Halpern & Pearl's own examples, **Def 2** does better than **Updated HP** and at least as good as **Modified HP**. Lastly we consider a very simple example that was offered as a counterexample to **Modified HP** by Rosenberg and Glymour [18].[27]

Example 6 We have equations $Y = X \vee D$ and $X = D$, and we consider a context such that $D = 1$. This looks very much like a standard case of overdetermination in which $X = 1$ and $D = 1$ are both overdetermining causes. That is also the verdict of all of the definitions considered in this paper, except for **Modified HP**: it does not consider $X = 1$ a cause of $Y = 1$. The reason for this is that $Y = 1$ depends counterfactually on $D = 1$ by itself, whereas it does not depend on $X = 1$ by itself and nor does it when we take $D = 1$ as a witness. Rosenberg and Glymour [18] state that Halpern endorses this conclusion, but offer the following story to motivate why they consider that an untenable position.

"An obedient gang is ordered by its leader to join him in murdering someone, and does so, all of them shooting the victim at the same time, or all of them together pushing the plunger connected to a bomb. The action of any one of the gang would suffice for the victim's death. If responsibility implies causality, whom among them is responsible? Were you among the jury, whom would you convict? What ought the Hague Court to do in cases of subordinates sure to obey orders? Halpern's theory says the gang leader and only the gang leader is a cause of the victim's death. This is a morally intolerable result; absent a plausible general principle severing responsibility from causation, any theory that yields such a result should be rejected."

Even if one disagrees with this judgment, the next section offers further motivation for preferring **Def 2** over **Modified HP**.

[25] A reviewer correctly pointed out that due to the complexity of the voting rule, it can be ambiguous to speak of *voting for or against outcome* j, because there exist contexts in which a voter can actually flip the outcome from $O = 1$ to $O = 0$ by changing their vote from 0 to 1. If we assume that the voters are unaware of the precise voting rule, we can ignore this complication.

[26] This is the formulation of the example found in Halpern [8, p. 109], but the example was first presented by Glymour et al. [4].

[27] I thank a reviewer for pointing out this example.

7.3 Def 2 vs the Others

Finally I will argue that **Def 2** does better than all of the other definitions on a few more examples according to two metrics: it offers verdicts that are both intuitively plausible *and* consistent across minor changes of the examples. Before doing so, I present an example that illustrates a special property of **Def 2**.

Recall from Section 3 that it is a necessary condition for all three HP definitions that there exists some $[\mathbf{W} \leftarrow \mathbf{w}]$ such that $Y = y$ counterfactually depends on $\mathbf{X} = \mathbf{x}$ under that intervention. The same is true for the most well-known definitions out there that have been inspired by the HP definitions (see Weslake [20] for an overview), as well as for **Def 3**, **Def 4**, and **Def 10**. Let us call definitions with this property *strongly counterfactual*. Although **Def 2** clearly also relies on counterfactuals, and thus falls within the counterfactual approach to causation, it is not strongly counterfactual, as the following example shows.[28]

Example 7 The equation for a binary variable Y is such that $Y = 1$ iff $N \neq 0$, and the range for N is $\{0, 1, 2, 3\}$. The equation for N is as follows: $N = 0$ if $A = 0$, $N = 1$ if $(A = 1 \wedge X = 1)$, $N = 2$ if $(A = 1 \wedge X = 0 \wedge W = 1)$, and $N = 3$ if $(A = 1 \wedge X = 0 \wedge W = 0)$. In a context where $A = W = X = 1$, we get that $X = 1$ causes $Y = 1$ according to **Def 2**. Yet there is no intervention such that $Y = 1$ depends on $X = 1$ under that intervention (and thus none of the other definitions would consider $X = 1$ a cause of $Y = 1$). In this case, both answers seem plausible. **Def 2** reaches its verdict because of the asymmetry between $(A = 1, X = 1)$ and $(A = 1, X = 0)$: only the former is by itself causally sufficient for a network that results in $Y = 1$, whereas the latter also needs the assistance of $W = 1$ or $W = 0$.

Now we consider six examples which are simple variations on the same theme, because they all share the following equation for Y: $Y = (X \wedge D) \vee A$. Moreover, they all share a context such that $X = 1$ and $A = 1$. The only difference between them lies with the value of D (0 or 1) and with the relation between A and D. (Concretely, there could be no relation, or it can be given by $A = D$, $A = \neg D$, $D = A$, and $D = \neg A$.) In all examples, all definitions agree that $A = 1$ is a cause of $Y = 1$. The disagreement arises over whether $X = 1$ should be considered a cause as well.

Intuitively, I would find it unacceptable to consider $X = 1$ a cause whenever $D = 0$, regardless of the relation between A and D. The disjunct in which X appears is false, and therefore it played no positive part whatsoever in causing $Y = 1$. Perhaps others are more tolerant. But even if that is the case, one should expect one's verdicts to exhibit some consistency. As we will see, **Def 2** and **Original HP** are the only definitions which can meet this demand.

The situation is simplest for **Original HP**: it considers $X = 1$ a cause of $Y = 1$ no matter what. To see why, take as a witness $(D = 1, A = 0)$. Holding fixed that witness, $Y = 1$ counterfactually depends on $X = 1$. Since $\mathbf{Z} = \{X\}$, the former is equivalent to AC2 for **Original HP**. So we gain consistency, but at the price of

[28] It is not so clear that **Def 8** also relies on counterfactuals, since it does not explicitly invoke counterfactual values of the candidate cause. Exploring this topic further lies beyond the scope of this paper.

extreme tolerance. In fact, Halpern and Pearl use precisely this example to argue against **Original HP** and in favor of **Updated HP** [10]:

Example 8 "Suppose that a prisoner dies either if X loads D's gun and D shoots, or if A loads and shoots his gun. Taking Y to represent the prisoner's death and making the obvious assumptions about the meaning of the variables, ... [we can use the equation described above]. Suppose that X loads D's gun ($X = 1$), D does not shoot ($D = 0$), but A does load and shoot his gun ($A = 1$), so that the prisoner dies. Clearly $A = 1$ is a cause of $Y = 1$. *We would not want to say that* $X = 1$ *is a cause of* $Y = 1$, *given that* D *did not shoot (i.e., given that* $D = 0$*).*" [emphasis added]

If we agree with Halpern and Pearl here – which I do – then **Original HP** can be discarded on the basis of this example (and on the basis of the many others we discussed previously, of course). I leave it to the reader to verify that none of the other definitions consider $X = 1$ to be a cause here.

However, the only definition that applies the intuition underlying this example to all cases in which $D = 0$ is **Def 2**. Moreover, it is the only remaining definition that offers a simple consistent answer in all cases: $X = 1$ is a cause of $Y = 1$ iff $D = 1$. To see why this is the case, we go over the possible directly sufficient sets. (Since X is only a parent of Y, we can invoke Proposition 6 and use **Def 3** instead of **Def 2**.) Clearly $X = 1$ is not directly sufficient for $Y = 1$ by itself. It is also clear that we cannot add $A = 1$ to the witness, because $A = 1$ is directly sufficient for $Y = 1$ all by itself. Therefore we are forced to choose D as our witness. If $D = 0$, this gives ($X = 1$, $D = 0$), which is not directly sufficient for $Y = 1$ and thus $X = 1$ is not a cause. If $D = 1$, we get ($X = 1$, $D = 1$), which is directly sufficient for $Y = 1$. Since the same does not hold for ($X = 0$, $D = 1$), $X = 1$ is a cause of $Y = 1$.

The following examples show that **Updated HP** and **Modified HP** flip-flop between calling $X = 1$ a cause or not even when holding fixed the value of D. Of course I cannot exclude the possibility that some consistent argumentation can be offered to explain the results of one of these definitions, but in its absence all of this speaks in favor of **Def 2**. We start with the three possible ways in which it can arise that $D = 1$.

Example 9 First consider the case where D is determined by the context, and we have a context such that $D = 1$. Here all four definitions agree that $X = 1$ is a cause of $Y = 1$.

Example 10 Second consider the case where the equation for D is given by $D = A$ and thus again $D = 1$ in the context under consideration. Here **Updated HP** and **Modified HP** flip their verdict, as they no longer consider $X = 1$ a cause of $Y = 1$.

Example 11 Third, we simply flip the relation between A and D so that $A = D$, and again $D = 1$ in the context under consideration. Now **Updated HP** and **Modified HP** go back to considering $X = 1$ a cause of $Y = 1$.

Next we consider the two remaining possible cases where $D = 0$ (Example 8 was the first such case).

Example 12 Consider the case where the equation for D is $D = \neg A$. As with Example 8, we have that $D = 0$, and yet **Updated HP** changes its verdict, calling $X = 1$ a cause of $Y = 1$.

Example 13 [29] Lastly, consider the case where the equation for D is $A = \neg D$, and thus we again have that $D = 0$. Now both **Modified HP** and **Updated HP** flip their verdicts as compared to Example 8. To see why, it suffices to consider **Modified HP**. The result for **Updated HP** then follows from Theorem 2. $D = 0$ by itself is not a cause of $Y = 1$ because there is no choice of witness that makes $Y = 1$ counterfactually depend on $D = 0$. Since $Y = 1$ does counterfactually depend on $(X = 1, D = 0)$, $X = 1$ is part of a cause of $Y = 1$.

8 Conclusion

I have developed twelve definitions of actual causation that formalize the NESS intuition with which Pearl started, and have shown that the most recent of the HP definitions is among them. Although these definitions vary widely in terms of the verdicts they reach, they all resemble each other as being instantiations of the same general definition. Each definition is made up of two elements: a definition of causal sufficiency, and a definition of necessity. Other definitions can easily be developed by playing around with these elements.

After studying various properties of these definitions and the relations between them, I moved on to the process of selecting the definition that does best in practice. In the majority of the many examples that we have considered, **Def 2** agrees with **Modified HP**. However, in Section 7.2 we came across two examples for which **Def 2** disagreed with **Modified HP** and where **Modified HP** gave the wrong verdict. Moreover, contrary to **Modified HP**, **Def 2** manages to give consistent (and intuitive) answers to the group of cases considered in the previous section. Therefore I conclude by suggesting that we should adopt **Def 2** as a definition of actual causation. This definition is made up of strong sufficiency and contrastive necessity. It states that $\mathbf{X} = \mathbf{x}$ causes $Y = y$ iff $\mathbf{X} = \mathbf{x}$ is a Minimal Contrastively Necessary Subset of a Strongly Sufficient Set for $Y = y$, or MCNS^4.

Appendix A: Causal Sufficiency

Proposition 1 $\mathbf{X} = \mathbf{x}$ *is strongly sufficient for* $\mathbf{Y} = \mathbf{y}$ *in* M *along a network* $\mathbf{N}$ *iff* $\mathbf{X} = \mathbf{x}$ *is strongly sufficient for* $\mathbf{Y} = \mathbf{y}$ *in* M.

Proof First assume $\mathbf{X} = \mathbf{x}$ is strongly sufficient for $\mathbf{Y} = \mathbf{y}$ in M and $\mathbf{N}$ can be used to show this. Then the result follows immediately from the observation that $\mathbf{X} = \mathbf{x}$

[29] The attentive reader will remember this example from the proof of Theorem 1.

is directly sufficient for $\mathbf{N} = \mathbf{n}$ and either $\mathbf{N} = \mathbf{n}$ is directly sufficient for $\mathbf{Y} = \mathbf{y}$ or $\mathbf{N} = \mathbf{Y}$ and $\mathbf{n} = \mathbf{y}$.

Second assume $\mathbf{X} = \mathbf{x}$ is strongly sufficient for $Y = y$ in M along a network $\mathbf{N}$. Define $\mathbf{A} = \mathcal{V} - (\mathbf{X} \cup \mathbf{N})$. We need to show that for all $\mathbf{a} \in \mathcal{R}(\mathbf{A})$ and all $\mathbf{u} \in \mathcal{R}(\mathcal{U})$ we have that $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}, \mathbf{A} \leftarrow \mathbf{a}]\mathbf{N} = \mathbf{n}$.

We know that $\mathbf{X} = \mathbf{x}$ is directly sufficient for $\mathbf{N_1} = \mathbf{n_1}$. Define $\mathbf{C_1} = \mathcal{V} - (\mathbf{X} \cup \mathbf{N_1})$ and $\mathbf{D_1} = \mathbf{N} - \mathbf{N_1}$. Note that $\mathbf{C_1} = \mathbf{A} \cup \mathbf{D_1}$. We have that for all $\mathbf{c_1} \in \mathcal{R}(\mathbf{C_1})$ and all $\mathbf{u} \in \mathcal{R}(\mathcal{U})$, $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}, \mathbf{C_1} \leftarrow \mathbf{c_1}]\mathbf{N_1} = \mathbf{n_1}$. In particular, we have that for all $\mathbf{a} \in \mathcal{R}(\mathbf{A})$ and all $\mathbf{u} \in \mathcal{R}(\mathcal{U})$, $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}, \mathbf{A} \leftarrow \mathbf{a}]\mathbf{N_1} = \mathbf{n_1}$.

Define $\mathbf{C_2} = \mathcal{V} - (\mathbf{N_1} \cup \mathbf{N_2})$ and $\mathbf{D_2} = \mathbf{N} - (\mathbf{N_1} \cup \mathbf{N_2})$. Note that $\mathbf{C_2} = \mathbf{A} \cup \mathbf{D_2} \cup \mathbf{X}$. We have that for all $\mathbf{c_2} \in \mathcal{R}(\mathbf{C_2})$ and all $\mathbf{u} \in \mathcal{R}(\mathcal{U})$, $(M, \mathbf{u}) \models [\mathbf{N_1} \leftarrow \mathbf{n_1}, \mathbf{C_2} \leftarrow \mathbf{c_2}]\mathbf{N_2} = \mathbf{n_2}$. In particular, we have that for all $\mathbf{a} \in \mathcal{R}(\mathbf{A})$ and all $\mathbf{u} \in \mathcal{R}(\mathcal{U})$, $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}, \mathbf{N_1} \leftarrow \mathbf{n_1}, \mathbf{A} \leftarrow \mathbf{a}]\mathbf{N_2} = \mathbf{n_2}$. Combined with the conclusion from the previous paragraph, it follows that for all $\mathbf{a} \in \mathcal{R}(\mathbf{A})$ and all $\mathbf{u} \in \mathcal{R}(\mathcal{U})$, $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}, \mathbf{A} \leftarrow \mathbf{a}]\mathbf{N_1} = \mathbf{n_1} \wedge \mathbf{N_2} = \mathbf{n_2}$.

Defining $\mathbf{N_{k+1}} = \mathbf{Y}$, we can generalize this reasoning for all consecutive $i \in \{3, \ldots, k+1\}$ to get the desired outcome. □

Appendix B: Defining Causation using Sufficiency

Theorem 1 *The following are all equivalences among the twelve definitions and the three HP definitions:*

- **Modified HP** *iff* **Def 1**
- **Def 2** *iff* **Def 5**
- **Def 8** *iff* **Def 11**
- **Def 3** *iff* **Def 6** *iff* **Def 9** *iff* **Def 12**

Proof First we consider the equivalences that do hold.

We start with the first equivalence: **Modified HP** iff **Def 1**. This is simply a matter of explicitly writing out the definitions, starting with actual weak sufficiency: $\mathbf{X} = \mathbf{x}$ is actually weakly sufficient for $Y = y$ in $(M, \mathbf{u})$ iff $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}]Y = y$. Next we note that the following condition is trivially satisfied for any $\mathbf{W} \subseteq \mathcal{V}$: $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}, \mathbf{W} \leftarrow \mathbf{w}^*]Y = y$.

Combining both claims, we can rewrite **Modified HP** as follows, which gives the desired result:

AC2(a). There is a set $\mathbf{W} \subseteq (\mathcal{V} - (\mathbf{X} \cup \{Y\}))$ and a setting $\mathbf{x}'$ of the variables in $\mathbf{X}$ such that $(\mathbf{X} = \mathbf{x}', \mathbf{W} = \mathbf{w}^*)$ is not actually weakly sufficient for $Y = y$ in $(M, \mathbf{u})$.

AC2(b). $(\mathbf{X} = \mathbf{x}, \mathbf{W} = \mathbf{w}^*)$ is actually weakly sufficient for $Y = y$ in $(M, \mathbf{u})$.

Next we consider all of the following equivalences: **Def 2** iff **Def 5**, **Def 8** iff **Def 11**, **Def 3** iff **Def 6**, **Def 9** iff **Def 12**. The reason we can group these together, is because we can prove all of them by invoking the following observation and two subsequent lemmas. □

Observation 1 *Recall our restriction on causal models that exogenous variables only appear in equations of the form $V = U$. Say $\mathbf{R} \subseteq \mathcal{V}$ are all variables which have such an equation, and call these the root variables. It is clear that if we intervene on all of the root variables, they take over the role of the exogenous variables. Concretely, given strong recursivity, for any setting $\mathbf{r} \in \mathcal{R}(\mathbf{R})$ there exists a unique setting $\mathbf{v} \in \mathcal{R}(\mathcal{V})$ so that for all contexts $\mathbf{u} \in \mathcal{R}(\mathcal{U})$ we have that $(M, \mathbf{u}) \models [\mathbf{R} \leftarrow \mathbf{r}]\mathcal{V} = \mathbf{v}$.*

Lemma 1 *Given a setting $\mathbf{X} = \mathbf{x}$, a setting $\mathbf{N} = \mathbf{n}$ that includes $Y = y$ and such that $\mathbf{N} \cap \mathbf{R} = \emptyset$, a context $\mathbf{u}$, the following holds:*[30]

- *$\mathbf{X} = \mathbf{x}$ is actually directly sufficient for $Y = y$ in $(M, \mathbf{u})$ iff $\mathbf{X} = \mathbf{x}$ is directly sufficient for $Y = y$ in M;*
- *$\mathbf{X} = \mathbf{x}$ is actually strongly sufficient for $Y = y$ in $(M, \mathbf{u})$ along $\mathbf{N} = \mathbf{n}$ iff $\mathbf{X} = \mathbf{x}$ is strongly sufficient for $Y = y$ in M along $\mathbf{N} = \mathbf{n}$.*

Proof Filling in the definitions of direct and actually direct sufficiency, the first equivalence reduces to the following: for all $\mathbf{c} \in \mathcal{R}(\mathcal{V} - (\mathbf{X} \cup \{Y\}))$, it holds that $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}, \mathbf{C} \leftarrow \mathbf{c}]Y = y$ iff for all $\mathbf{u}'' \in \mathcal{R}(\mathcal{U})$, $(M, \mathbf{u}'') \models [\mathbf{X} \leftarrow \mathbf{x}, \mathbf{C} \leftarrow \mathbf{c}]Y = y$.

Because of Observation 1, we have that for any setting $\mathbf{v} \in \mathcal{V}$ and any setting $\mathbf{r} \in \mathcal{R}(\mathbf{R})$, it holds that $(M, \mathbf{u}) \models [\mathbf{R} \leftarrow \mathbf{r}]\mathcal{V} = \mathbf{v}$ iff for all contexts $\mathbf{u}'' \in \mathcal{R}(\mathcal{U})$, $(M, \mathbf{u}'') \models [\mathbf{R} \leftarrow \mathbf{r}]\mathcal{V} = \mathbf{v}$. Combining this with the fact that $\mathbf{R} \subseteq (\mathbf{C} \cup \mathbf{X})$ gives the desired result.

The second equivalence can be reformulated as follows: $\mathbf{X} = \mathbf{x}$ is actually directly sufficient for $\mathbf{N} = \mathbf{n}$ in $(M, \mathbf{u})$ iff $\mathbf{X} = \mathbf{x}$ is directly sufficient for $\mathbf{N} = \mathbf{n}$ in M. In turn, this reduces to: for all $\mathbf{c} \in \mathcal{R}(\mathcal{V} - (\mathbf{X} \cup \mathbf{N}))$, it holds that $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}, \mathbf{C} \leftarrow \mathbf{c}]\mathbf{N} = \mathbf{n}$ iff for all $\mathbf{u}'' \in \mathcal{R}(\mathcal{U})$, $(M, \mathbf{u}'') \models [\mathbf{X} \leftarrow \mathbf{x}, \mathbf{C} \leftarrow \mathbf{c}]\mathbf{N} = \mathbf{n}$.

Given that $\mathbf{N} \cap \mathbf{R} = \emptyset$, we still have that $\mathbf{R} \subseteq (\mathbf{C} \cup \mathbf{X})$, and therefore we can apply the same reasoning as before. □

Lemma 2 *For all twelve instances of the **General Definition of Causation** we can restrict ourselves to sets $\mathbf{N}$ so that $(\mathbf{N} - \{Y\}) \cap \mathbf{R} = \emptyset$.*

Proof Let $\mathbf{A}$ denote $(\mathbf{N} - \{Y\}) \cap \mathbf{R}$. For all definitions using either variants of direct or weak sufficiency the result follows immediately from the fact that $\mathbf{N} - \{Y\} = \emptyset$.

First consider the case where we use non-actual strong sufficiency (**Def 5** or **Def 11**). In that case, AC2(b) can never be satisfied unless $\mathbf{A} = \emptyset$. To see why, note that in all contexts $\mathbf{u}'' \in \mathcal{R}(\mathcal{U})$, it has to hold that $(M, \mathbf{u}'') \models [\mathbf{X} \leftarrow \mathbf{x}, \mathbf{W} \leftarrow \mathbf{w}^*]\mathbf{A} = \mathbf{a}$. Since $\mathbf{A} \cap (\mathbf{X} \cup \mathbf{W})$ and the equation for each element $A_i \in \mathbf{A}$ is of the form $A_i = U$ for some exogenous variable U, this is impossible. (Strictly speaking it is possible, namely if the range of U consists only of the single value a_i^*. Although I did not make this explicit in Section 2, it is standard to assume that all variables have a range that contains at least two elements.)

[30] $\mathbf{R}$ is defined in Observation 1.

Second consider the case where we use actual strong sufficiency and contrastive necessity (**Def 2**). (The case of **Def 8** is entirely analogous.) Say we are considering a candidate cause $\mathbf{X} = \mathbf{x}$, a candidate witness $\mathbf{W} = \mathbf{w}^*$, contrast values $\mathbf{x}'$, and a setting $\mathbf{N} = \mathbf{n}$ that includes $Y = y$. Given AC1, we can safely assume that $\mathbf{n} = \mathbf{n}^*$.

I claim that the following holds, from which the result follows: $\mathbf{X} = \mathbf{x}$ satisfies AC2 using contrast values $\mathbf{x}'$, witness $\mathbf{W} = \mathbf{w}^*$, and network $\mathbf{N}$ iff $\mathbf{X} = \mathbf{x}$ satisfies AC2 using contrast values $\mathbf{x}'$, witness $(\mathbf{W} = \mathbf{w}^*, \mathbf{A} = \mathbf{a}^*)$, and network $\mathbf{N} - \mathbf{A}$.

Because $\mathbf{A} \subseteq \mathbf{R}$, we have that for any set $\mathbf{B} \subseteq (\mathcal{V} - \mathbf{A})$, and any setting $\mathbf{b} \in \mathcal{R}(\mathbf{B})$, $(M, \mathbf{u}) \models [\mathbf{B} \leftarrow \mathbf{b}]\mathbf{A} = \mathbf{a}^*$. Moreover, since $(M, \mathbf{u}) \models \mathbf{A} = \mathbf{a}^*$, for each setting $\mathbf{v} \in (\mathcal{V} - \mathbf{A})$ we also have that $(M, \mathbf{u}) \models [\mathbf{B} \leftarrow \mathbf{b}](\mathcal{V} - \mathbf{A}) = \mathbf{v}$ iff $(M, \mathbf{u}) \models [\mathbf{B} \leftarrow \mathbf{b}, \mathbf{A} \leftarrow \mathbf{a}^*](\mathcal{V} - \mathbf{A}) = \mathbf{v}$.

Using these observations and the fact that $\mathbf{A} \subseteq \mathbf{N}$, we get that the following two conditions are equivalent, for which the result follows as far as AC2(b) is concerned:

AC2(b). For all $\mathbf{c} \in \mathcal{R}(\mathcal{V} - (\mathbf{X} \cup \mathbf{W} \cup \mathbf{N}))$ we have that $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}, \mathbf{W} \leftarrow \mathbf{w}^*, \mathbf{C} \leftarrow \mathbf{c}]\mathbf{N} = \mathbf{n}^*$.

AC2(b). For all $\mathbf{c} \in \mathcal{R}(\mathcal{V} - (\mathbf{X} \cup \mathbf{W} \cup \mathbf{N}))$ we have that $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}, \mathbf{W} \leftarrow \mathbf{w}^*, \mathbf{A} \leftarrow \mathbf{a}^*, \mathbf{C} \leftarrow \mathbf{c}](\mathbf{N} - \mathbf{A}) = \mathbf{n_2}^*$ (where $\mathbf{n_2}$ is the restriction of $\mathbf{n}^*$ to $(\mathbf{N} - \mathbf{A})$).

Now we focus on AC2(a^c).

Let us first assume AC2(a^c) holds for $\mathbf{X} = \mathbf{x}$, contrast values $\mathbf{x}'$, witness $(\mathbf{W} = \mathbf{w}^*, \mathbf{A} = \mathbf{a}^*)$, and network $\mathbf{N} - \mathbf{A}$. We need to show that it holds for $\mathbf{X} = \mathbf{x}$, contrast values $\mathbf{x}'$, witness $(\mathbf{W} = \mathbf{w}^*)$, and network $\mathbf{N}$.

Consider some $\mathbf{S} \subseteq \mathbf{N}$ with $Y \in \mathbf{S}$. We need to find a $\mathbf{t} \in \mathcal{R}(\mathcal{V} - (\mathbf{X} \cup \mathbf{W} \cup \mathbf{S}))$ so that $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}', \mathbf{W} \leftarrow \mathbf{w}^*, \mathbf{T} \leftarrow \mathbf{t}]\mathbf{S} \neq \mathbf{s}^*$. Define $\mathbf{S_1} = \mathbf{S} - \mathbf{A}$, $\mathbf{S_2} = \mathbf{S} \cap \mathbf{A}$, and $\mathbf{A_1} = \mathbf{A} - \mathbf{S}$.

Since $\mathbf{S_1} \subseteq (\mathbf{N} - \mathbf{A})$ with $Y \in \mathbf{S_1}$, we know that there exists some $\mathbf{t_1} \in \mathcal{R}(\mathcal{V} - (\mathbf{X} \cup \mathbf{W} \cup \mathbf{A} \cup \mathbf{S_1})$ so that $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}', \mathbf{W} \leftarrow \mathbf{w}^*, \mathbf{A} \leftarrow \mathbf{a}^*, \mathbf{T} \leftarrow \mathbf{t_1}]\mathbf{S_1} \neq \mathbf{s_1}^*$. Since $\mathbf{S_1} \subseteq \mathbf{S}$, it also holds that $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}', \mathbf{W} \leftarrow \mathbf{w}^*, \mathbf{A} \leftarrow \mathbf{a}^*, \mathbf{T} \leftarrow \mathbf{t_1}]\mathbf{S} \neq \mathbf{s}^*$. Also, given our observations about $\mathbf{A}$, it also follows that $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}', \mathbf{W} \leftarrow \mathbf{w}^*, \mathbf{A_1} \leftarrow \mathbf{a_1}, \mathbf{T} \leftarrow \mathbf{t_1}]\mathbf{S} \neq \mathbf{s}^*$. Lastly, note that $[\mathcal{V} - (\mathbf{X} \cup \mathbf{W} \cup \mathbf{A} \cup \mathbf{S_1})] \cup \mathbf{A_1} = \mathcal{V} - (\mathbf{X} \cup \mathbf{W} \cup \mathbf{S})$. Therefore we can choose $\mathbf{t} = (\mathbf{a_1}, \mathbf{t_1})$.

Next we consider the other direction: assume AC2(a^c) holds for $\mathbf{X} = \mathbf{x}$, contrast values $\mathbf{x}'$, witness $\mathbf{W} = \mathbf{w}^*$, and network $\mathbf{N}$. We need to show that it holds for $\mathbf{X} = \mathbf{x}$, contrast values $\mathbf{x}'$, witness $(\mathbf{W} = \mathbf{w}^*, \mathbf{A} = \mathbf{a}^*)$, and network $\mathbf{N} - \mathbf{A}$.

Consider some $\mathbf{S} \subseteq (\mathbf{N} - \mathbf{A})$ with $Y \in \mathbf{S}$. We need to find a $\mathbf{t} \in \mathcal{R}(\mathcal{V} - (\mathbf{X} \cup \mathbf{W} \cup \mathbf{A} \cup \mathbf{S})$ so that $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}', \mathbf{W} \leftarrow \mathbf{w}^*, \mathbf{A} \leftarrow \mathbf{a}^*, \mathbf{T} \leftarrow \mathbf{t}]\mathbf{S} \neq \mathbf{s}^*$.

Note that $(\mathbf{S} \cup \mathbf{A}) \subseteq \mathbf{N}$, and also $Y \in (\mathbf{S} \cup \mathbf{A})$. Therefore there exists some $\mathbf{t_2} \in \mathcal{R}(\mathcal{V} - (\mathbf{X} \cup \mathbf{W} \cup \mathbf{A} \cup \mathbf{S})$ so that $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}', \mathbf{W} \leftarrow \mathbf{w}^*, \mathbf{A} \leftarrow \mathbf{a}^*, \mathbf{T} \leftarrow \mathbf{t_2}](\mathbf{S} \neq \mathbf{s}^* \vee \mathbf{A} \neq \mathbf{a}^*)$. It follows that $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}', \mathbf{W} \leftarrow \mathbf{w}^*, \mathbf{A} \leftarrow \mathbf{a}^*, \mathbf{T} \leftarrow \mathbf{t_2}]\mathbf{S} \neq \mathbf{s}^*$. Choosing $\mathbf{t} = \mathbf{t_2}$ gives the desired result. □

Because of the above lemmas, all that remains is to show that the above equivalences hold also when $Y \in \mathbf{R}$. This is accomplished by showing that settings of such variables do not have any cause, regardless of the definition one uses.

AC2(a) requires us to look at all subsets of $\mathbf{N} = \mathbf{n}$ that include $Y = y$, and verify that the candidate cause and witness $(\mathbf{X} = \mathbf{x}', \mathbf{W} = \mathbf{w}^*)$ (or candidate witness $\mathbf{W} = \mathbf{w}^*$ in case we use AC2(a^m)) is not sufficient for that subset. One such subset is the one containing just $Y = y$. By AC1, we have that $(M, \mathbf{u}) \models Y = y$. Since $Y \in \mathbf{R}$, there is no intervention on the other endogenous variables so that $Y \neq y$ under that intervention in $\mathbf{u}$. Therefore any definition of causation using a version of *actual* sufficiency (i.e., **Def 2**, **Def 3**, **Def 8**, and **Def 9**) considers all sets that do not include Y to be sufficient for $Y = y$ in $(M, \mathbf{u})$. In particular, they consider $(\mathbf{X} = \mathbf{x}', \mathbf{W} = \mathbf{w}^*)$ to be sufficient for $Y = y$ in $(M, \mathbf{u})$, and thus fail to meet condition AC2(a).

For the definitions using *non-actual* variants of sufficiency (**Def 5**, **Def 6**, **Def 11**, and **Def 12**), it is condition AC2(b) that can never be satisfied. Analogous to what we saw in the proof of Lemma 2, this follows from the fact that whatever version of sufficiency we use, $Y = y$ has to hold in all contexts, which is impossible given that $Y \notin (\mathbf{X} \cup \mathbf{W})$. From this the result follows.

Now we prove the only remaining equivalence: **Def 6** iff **Def 12**. (Given the previous equivalences, other choices are possible too.) We need to show that the following two statements are equivalent:

- $\mathbf{W} = \mathbf{w}^*$ is not directly sufficient for $Y = y$.
- There exists values $\mathbf{x}'$ of $\mathbf{X}$ such that $(\mathbf{X} = \mathbf{x}', \mathbf{W} = \mathbf{w}^*)$ is not directly sufficient for $Y = y$.

Filling in Definition 7, the result follows immediately:

- There exists a $\mathbf{z} \in \mathcal{R}(\mathcal{V} - (\mathbf{W} \cup \mathbf{X} \cup \{Y\}))$, a $\mathbf{x}' \in \mathcal{R}(\mathbf{X})$, and a $\mathbf{u}' \in \mathcal{R}(\mathcal{U})$ so that $(M, \mathbf{u}') \models [\mathbf{W} \leftarrow \mathbf{w}^*, \mathbf{X} \leftarrow \mathbf{x}', \mathbf{C} \leftarrow \mathbf{c}]Y \neq y$.
- There exists values $\mathbf{x}'$ of $\mathbf{X}$, a $\mathbf{z} \in \mathcal{R}(\mathcal{V} - (\mathbf{W} \cup \mathbf{X} \cup \{Y\}))$ and a $\mathbf{u}' \in \mathcal{R}(\mathcal{U})$ so that $(M, \mathbf{u}') \models [\mathbf{W} \leftarrow \mathbf{w}^*, \mathbf{X} \leftarrow \mathbf{x}', \mathbf{C} \leftarrow \mathbf{c}]Y \neq y$.

Second, we go over some examples to show that none of the other equivalences hold. (Obviously, from now on we may ignore **Def 1**, **Def 5**, **Def 6**, **Def 7**, **Def 9**, **Def 11**, and **Def 12**.)

Example 14 Equations: $Y = (X \wedge A) \vee D$, $D = A$. Context: $A = 1$. Then $X = 1$ is a cause of $Y = 1$ according to:

- **Modified HP**: We can always consider choosing $\mathbf{W} = \emptyset$, in which case we simply get counterfactual dependence: $(M, \mathbf{u}) \models \mathbf{X} = \mathbf{x} \wedge Y = 1$ and $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}']Y \neq y$. Doing so in this example, we see that $Y = 1$ counterfactually depends on $(X = 1, D = 1)$. There is clearly also no witness $\mathbf{W} = \mathbf{w}^*$ to show that $X = 1$ or $D = 1$ are causes by themselves, so $X = 1$ is part of a cause.
- **Updated HP** and **Original HP**: taking $(A = 1, D = 0)$ as a witness meets the conditions.
- **Def 3**: again take $(A = 1, D = 0)$ as a witness.
- **Def 2**: follows from the previous item and Theorem 4.
- **Def 8**: follows from the previous item and Theorem 4.

$X = 1$ is not a cause of $Y = 1$ according to:

- **Def 10**: $X = 1$ by itself does not weakly suffice for $Y = 1$ (just look at a context in which $A = 0$), so we need to add A or D to the witness. But both $A = 1$ and $D = 1$ each weakly suffice for $Y = 1$.
- **Def 4**: $(X = 0, A = 1)$ and $(X = 0, D = 1)$ also weakly suffice for $Y = 1$.

So we know that **Def 4** and **Def 10** are not equivalent to any of the other definitions. We give an example to show that **Def 4** and **Def 10** are not equivalent to each other either.

Example 15 Equations: $Y = X \wedge A$, $X = A$. Context: $A = 1$. Since $X = 1$ is not weakly sufficient for $Y = 1$, we need to include $A = 1$ in the witness. Indeed, $(X = 1, A = 1)$ is weakly sufficient for $Y = 1$. However, so is $A = 1$, and therefore $X = 1$ does not cause $Y = 1$ according to **Def 10**. Yet $(X = 0, A = 1)$ is not weakly sufficient for $Y = 1$, and therefore $X = 1$ causes $Y = 1$ according to **Def 4**.

This leaves us with the HP definitions, **Def 2**, **Def 3**, and **Def 8**. The next example shows that the former are not equivalent to the latter.

Example 16 Equations: $Y = (X \wedge \neg A) \vee D$, $D = A$. Context: $A = 1$. Then $X = 1$ is a cause of $Y = 1$ according to:

- **Modified HP**: $Y = 1$ counterfactually depends on $(X = 1, A = 1)$, and not on either $X = 1$ or $A = 1$. So $X = 1$ is part of a cause.
- **Updated HP** and **Original**: take $A = 0$ as a witness.

$X = 1$ is not a cause of $Y = 1$ according to:

- **Def 3**: $X = 1$ by itself does not directly suffice for $Y = 1$ (just look at $[A \leftarrow 1, D \leftarrow 0]$), so we need to add A or D to the witness. Since the actual value of A is 1, it is of no use, which leaves us with D. But $D = 1$ directly suffices for $Y = 1$ by itself, and thus so does $(X = 0, D = 1)$.
- **Def 2**: follows from the previous item and Proposition 12.
- **Def 8**: follows from the previous item and Proposition 12.

That none of the HP definitions are equivalent is of course a well-established fact, and also follows from the examples we consider in Section 7. Therefore we are left with showing that **Def 2**, **Def 3**, and **Def 8** are not equivalent. That **Def 3** differs from the other two is a direct consequence of some of our later results, but a simple example illustrates this as well.

Example 17 Equations: $Y = A$, $A = X$. Context: $A = 1$. Then it is easy to see that $X = 1$ causes $Y = 1$ according to all definitions here considered, except for **Def 3**.

Lastly, I refer the reader to Example 2 in Sections 7 for an example that shows **Def 2** and **Def 8** are not equivalent.

Proposition 4 *If* $\mathbf{X} = \mathbf{x}$ *causes* $Y = y$ *in* $(M, \mathbf{u})$ *according to a definition that uses minimal necessity, then* $\mathbf{X}$ *is a singleton.*

Proof Since we know that **Def 7** is unsatisfiable and we have Theorem 3, we only need to consider **Def 3**, **Def 8**, and **Def 10**. The following applies to both weak and direct sufficiency (i.e., **Def 3** and **Def 10**.)

Assume $(\mathbf{X_1} = \mathbf{x_1}, \mathbf{X_2} = \mathbf{x_2}, \mathbf{W} = \mathbf{w}^*)$ is sufficient for $Y = y$, and $\mathbf{W} = \mathbf{w}^*$ is not sufficient for $Y = y$. If either $(\mathbf{X_2} = \mathbf{x_2}, \mathbf{W} = \mathbf{w}^*)$ or $(\mathbf{X_1} = \mathbf{x_1}, \mathbf{W} = \mathbf{w}^*)$ is also sufficient for $Y = y$, then $(\mathbf{X_1} = \mathbf{x_1}, \mathbf{X_2} = \mathbf{x_2})$ is not minimal.

So let us assume that neither $(\mathbf{X_2} = \mathbf{x_2}, \mathbf{W} = \mathbf{w}^*)$ nor $(\mathbf{X_1} = \mathbf{x_1}, \mathbf{W} = \mathbf{w}^*)$ is sufficient for $Y = y$. This means we can move $\mathbf{X_2}$ to the witness to show that $\mathbf{X_1} = \mathbf{x_1}$ satisfies AC2 by itself, and likewise for $\mathbf{X_2}$ and $\mathbf{X_1}$ reversed. From this the result follows.

Now we prove that it also holds for strong sufficiency, i.e., for **Def 8**. Assume $(\mathbf{X_1} = \mathbf{x_1}, \mathbf{X_2} = \mathbf{x_2}, \mathbf{W} = \mathbf{w}^*)$ is sufficient for $Y = y$ along $\mathbf{N}$, and $\mathbf{W} = \mathbf{w}^*$ is not sufficient for $Y = y$ along any network $\mathbf{S} \subseteq \mathbf{N}$. If either $(\mathbf{X_2} = \mathbf{x_2}, \mathbf{W} = \mathbf{w}^*)$ or $(\mathbf{X_1} = \mathbf{x_1}, \mathbf{W} = \mathbf{w}^*)$ is also sufficient for $Y = y$ along $\mathbf{N}$, then $(\mathbf{X_1} = \mathbf{x_1}, \mathbf{X_2} = \mathbf{x_2})$ is not minimal.

So let us assume that neither $(\mathbf{X_2} = \mathbf{x_2}, \mathbf{W} = \mathbf{w}^*)$ nor $(\mathbf{X_1} = \mathbf{x_1}, \mathbf{W} = \mathbf{w}^*)$ is sufficient for $Y = y$ along $\mathbf{N}$. If the same is true for all subnetworks $\mathbf{S} \subseteq \mathbf{N}$, then as before, we can move either one of $\mathbf{X_1}$ and $\mathbf{X_2}$ to the witness to show that the other satisfies AC2 by itself.

So let us assume that there is some subnetwork $\mathbf{S}' \subseteq \mathbf{N}$ such that $(\mathbf{X_1} = \mathbf{x_1}, \mathbf{W} = \mathbf{w}^*)$ is sufficient for $Y = y$ along $\mathbf{S}'$. (Obviously the same reasoning applies to $\mathbf{X_2}$.) Since all subnetworks $\mathbf{S}''$ of $\mathbf{S}'$ are also subnetworks of $\mathbf{N}$, it follows from the above that $(\mathbf{X_1} = \mathbf{x_1})$ satisfies AC2 by itself when taking $\mathbf{W}$ as witness and $\mathbf{S}'$ as network. From this the result follows. □

Theorem 2 *The only implications – involving either causes or parts of causes – between the remaining five definitions* (**Def 2**, **Def 3**, **Def 4**, **Def 8**, *and* **Def 10**) *and the three HP definitions are the following ones (and their immediate consequences, of course):*

- *If part of* **Modified HP** *then* **Updated HP**;
- *If part of* **Updated HP** *then* **Original HP**;
- *If* **Def 3** *then* **Def 2**;
- *If part of* **Def 2** *then* **Def 8**;
- *If* **Def 3** *then* **Original HP**;
- *If* **Def 10** *then* **Def 4**.

Proof The first two implications are proven in Halpern [8].

First we prove the third implication. Assume $\mathbf{X} = \mathbf{x}$ causes $Y = y$ with witness $\mathbf{W}$ according to **Def 3**. It follows from Proposition 10 that $\mathbf{X}$ is a single conjunct X. Note that this immediately implies minimality of $\mathbf{X}$.

In other words, $(X = x, \mathbf{W} = \mathbf{w}^*)$ is directly sufficient for $Y = y$, and there exists some x' such that $(X = x', \mathbf{W} = \mathbf{w}^*)$ is not directly sufficient for $Y = y$. From the former it follows that $(X = x, \mathbf{W} = \mathbf{w}^*)$ is strongly sufficient for $Y = y$ along $\emptyset$. From the latter it follows that $(X = x', \mathbf{W} = \mathbf{w}^*)$ is not strongly sufficient for $Y = y$ along $\emptyset$, from which the result follows.

Second we prove the fourth implication. Assume $(X = x, \mathbf{X_2} = \mathbf{x_2}, \mathbf{W} = \mathbf{w}^*)$ is sufficient for $Y = y$ along $\mathbf{N}$, and $(X = x', \mathbf{X_2} = \mathbf{x_2}', \mathbf{W} = \mathbf{w}^*)$ is not sufficient for $Y = y$ along any network $\mathbf{S} \subseteq \mathbf{N}$, for some $\mathbf{N}$, x' and $\mathbf{x_2}'$. We show that $X = x$ causes $Y = y$ according to **Def 8**.

Taking $(\mathbf{X_2} = \mathbf{x_2}, \mathbf{W} = \mathbf{w}^*)$ as our witness and using $\mathbf{N}$, AC2(b) remains unchanged. If $(\mathbf{X_2} = \mathbf{x_2}, \mathbf{W} = \mathbf{w}^*)$ is not sufficient for $Y = y$ along any network $\mathbf{S} \subseteq \mathbf{N}$, then the result follows. We proceed by a reductio.

Let us assume that $(\mathbf{X_2} = \mathbf{x_2}, \mathbf{W} = \mathbf{w}^*)$ is sufficient for $Y = y$ along some $\mathbf{S} \subseteq \mathbf{N}$. If $(\mathbf{X_2} = \mathbf{x_2}', \mathbf{W} = \mathbf{w}^*)$ is not sufficient for $Y = y$ along any $\mathbf{S}'' \subseteq \mathbf{S}$, we have a violation of minimality (since X is redundant). Therefore we know that $(\mathbf{X_2} = \mathbf{x_2}', \mathbf{W} = \mathbf{w}^*)$ is sufficient for $Y = y$ along some network $\mathbf{S}'' \subseteq \mathbf{S}$.

This means that there exist values $\mathbf{s}'' \in \mathcal{R}(\mathbf{S}'')$ so that for all settings $\mathbf{c} \in \mathcal{R}(\mathcal{V} - (\mathbf{S}'' \cup \mathbf{X_2} \cup \{X, Y\})$, and for all $x'' \in \mathcal{R}(X)$, it holds that $(M, \mathbf{u}) \models [\mathbf{X_2} \leftarrow \mathbf{x_2}', \mathbf{W} \leftarrow \mathbf{w}^*, \mathbf{C} \leftarrow \mathbf{c}, X \leftarrow x'']\mathbf{S} = \mathbf{s}''$ and $(M, \mathbf{u}) \models [\mathbf{X_2} \leftarrow \mathbf{x_2}', \mathbf{W} \leftarrow \mathbf{w}^*, \mathbf{C} \leftarrow \mathbf{c}, X \leftarrow x'', \mathbf{S} \leftarrow \mathbf{s}'']Y = y$. In particular, this holds if we choose $X = x'$. But that means that $(X = x', \mathbf{X_2} = \mathbf{x_2}', \mathbf{W} = \mathbf{w}^*)$ is also sufficient for $Y = y$ along $\mathbf{S}''$, which contradicts our starting assumption.

Third we prove the fifth implication. As with the third implication, assume that $(X = x, \mathbf{W} = \mathbf{w}^*)$ is directly sufficient for $Y = y$, and there exists some x' such that $(X = x', \mathbf{W} = \mathbf{w}^*)$ is not directly sufficient for $Y = y$. From the latter it follows that there exists a setting $\mathbf{d}$ of $\mathcal{V} - (\mathbf{X} \cup \mathbf{W} \cup \{Y\})$ such that $(M, \mathbf{u}) \models [X \leftarrow x', \mathbf{W} \leftarrow \mathbf{w}^*, \mathbf{D} \leftarrow \mathbf{d}]Y \neq y$. This means that if we take $(\mathbf{W} = \mathbf{w}^*, \mathbf{D} = \mathbf{d})$ as witness, AC2(a) is satisfied for **Original HP**. Since $(X = x, \mathbf{W} = \mathbf{w}^*)$ is directly sufficient for $Y = y$, we know that $(M, \mathbf{u}) \models [X \leftarrow x, \mathbf{W} \leftarrow \mathbf{w}^*, \mathbf{D} \leftarrow \mathbf{d}]Y = y$. Also, we have that $\mathbf{Z} = \mathbf{X}$, and thus the former means that also AC2(b) is satisfied for **Original HP**.

Fourth we prove the last implication. Assume $X = x$ causes $Y = y$ with witness $\mathbf{W}$ according to **Def 10**. (We know because of Proposition 10 that $\mathbf{X}$ is a singleton.) In other words, $(X = x, \mathbf{W} = \mathbf{w}^*)$ is weakly sufficient for $Y = y$, and $\mathbf{W} = \mathbf{w}^*$ is not weakly sufficient for $Y = y$. Remains to be shown that there exist a value x' so that $(X = x', \mathbf{W} = \mathbf{w}^*)$ is not weakly sufficient for $Y = y$.

Say $\mathbf{u}'$ is a context such that $(M, \mathbf{u}') \models [\mathbf{W} \leftarrow \mathbf{w}^*]Y \neq y$, and say x' is the unique value such that $(M, \mathbf{u}') \models [\mathbf{W} \leftarrow \mathbf{w}^*]X = x'$. Then also $(M, \mathbf{u}') \models [X \leftarrow x', \mathbf{W} \leftarrow \mathbf{w}^*]Y \neq y$, which is what remained to be shown.

Fifth, we show that none of the remaining implications hold. (Again, we do not consider the relations amongst the HP definitions explicitly and refer the reader to the examples in Section 7. We also do not explicitly consider the remaining implications for parts of causes, but the reader can verify that the following examples suffice to falsify all those implications as well. For the left-hand side of all implications this follows immediately from the fact that the causes in all the following examples are singletons. For the right-hand side of implications, Propositions 4, 5, and 6 come in handy.)

Example 15 shows that **Def 4** does not imply **Def 10**.

Example 14 shows that none of the other definitons imply either **Def 4** or **Def 10**. So there are no remaining implications with either **Def 4** or **Def 10** on the right-hand side.

Example 17 shows that **Def 3** is not implied by any definition.

Example 16 shows that none of the HP definitions imply **Def 2** or **Def 8**. Note that **Def 4** and **Def 10** also consider $X = 1$ a cause of $Y = 1$ in that example (since $X = 1$ is weakly sufficient for $Y = 1$, whereas $X = 0$ or the emptyset is not). Further, Example 2 shows that **Def 8** does not imply **Def 2**. Therefore there are no remaining implications with **Def 2** or **Def 8** on the right-hand side.

That leaves us to consider implications with one of the HP definitions on the right-hand side. Given the first two implications of Theorem 4, it suffices to show that none of **Def 4**, **Def 2**, **Def 8**, or **Def 10**, imply **Original HP**, and that **Def 3** does not imply **Updated HP**.

I refer the reader to Example 7 in Section 7 for an example where **Def 2** – and thus also **Def 8** – hold and **Original HP** does not. □

The following example shows that neither **Def 4** nor **Def 10** implies **Original HP**.

Example 18 Equations: $Y = Z_1 \vee Z_2 \vee A$, $Z_1 = X \wedge A$, $Z_2 = X \wedge \neg A$. Context: $A = 1$ and $X = 1$. Then $X = 1$ is a cause of $Y = 1$ according to:

- **Def 10**: $X = 1$ is weakly sufficient for $Y = 1$ and $\emptyset$ is not.
- **Def 4**: follows from the previous one.

Yet $X = 1$ is not a cause of $Y = 1$ according to **Original HP**. To see why, note that we need to include $A = 0$ into the witness in order to get AC2(a), and we must exclude Z_1. Also, we clearly cannot add $Z_2 = 1$. Therefore the witness has to be $A = 0$. The actual value of Z_2 is 0. Since we have $(M, \mathbf{u}) \models [X \leftarrow 1, A \leftarrow 0, Z_2 \leftarrow 0]Y = 0$, AC2(b) is not satisfied.

Lastly, an example to show that **Def 3** does not imply **Updated HP**.

Example 19 Equations: $Y = (X \wedge D) \vee A$, $D = A$. Context: $A = 1$ and $X = 1$. Then $X = 1$ is a cause of $Y = 1$ according to **Def 3**: $(X = 1, D = 1)$ is directly sufficient for $Y = 1$, and $(X = 0, D = 1)$ is not. But $X = 1$ is not a cause of $Y = 1$ according to **Updated HP**. To see why, note that we need to include $A = 0$ into the witness in order to get AC2(a). But $(M, \mathbf{u}) \models [X \leftarrow 1, A \leftarrow 0]Y = 0$, thus falsifying AC2(b) for **Updated HP**.

Appendix C: Excluding Def 3 and Def 10

Proposition 5 *If* $\mathbf{X} = \mathbf{x}$ *causes* $Y = y$ *in* $(M, \mathbf{u})$ *according to* ***Def 3****, then* $\mathbf{X}$ *is a singleton, and* X *is a parent of* Y.

Proof That $\mathbf{X}$ is always a singleton is a direct consequence of the combination of Proposition 10 and Theorem 3.

Recall that X is a parent of Y iff there exists a context $\mathbf{u}''$, a setting $\mathbf{z} \in \mathcal{R}(\mathcal{V} - \{X, Y\})$, and values x, x'' of X so that $F_Y(\mathbf{u}'', \mathbf{z}, x) \neq F_Y(\mathbf{u}'', \mathbf{z}, x'')$. This means precisely that for some $y \in \mathcal{R}(Y)$, $(M, \mathbf{u}'') \models [\mathbf{Z} \leftarrow \mathbf{z}, X \leftarrow x]Y = y$ and

$(M, \mathbf{u}'') \models [\mathbf{Z} \leftarrow \mathbf{z}, X \leftarrow x'']Y \neq y$. If $X = x$ causes $Y = y$ according to **Def 3**, the existence of values such that the previous holds follows immediately. □

Proposition 6 *If X is only a parent of Y, then* **Def 3**, **Def 2**, *and* **Def 8** *are all equivalent for causes $X = x$.*

Proof Given Theorem 4, we only need to prove the implication from **Def 8** to **Def 3**.

Assume X is only a parent of Y, and $X = x$ causes $Y = y$ according to **Def 8**. Thus, there is a witness $\mathbf{W}$ and some network $\mathbf{N}$ such that $(X = x, \mathbf{W} = \mathbf{w}^*)$ is strongly sufficient for $Y = y$ along $\mathbf{N}$, and $(\mathbf{W} = \mathbf{w}^*)$ is not strongly sufficient for $Y = y$ along any subnetwork of $\mathbf{N}$.

First consider the case where $\mathbf{N} = \emptyset$. This means that $(X = x, \mathbf{W} = \mathbf{w}^*)$ is directly sufficient for $Y = y$, and $(\mathbf{W} = \mathbf{w}^*)$ is not directly sufficient for $Y = y$. That means precisely that $X = x$ causes $Y = y$ according to **Def 12**. The result now follows from Theorem 3.

Second consider the case where there exists some $N \in \mathbf{N}$. If N is not an ancestor of Y, it can be removed from $\mathbf{N}$ without consequence. If N is an ancestor of Y, then it cannot be a descendant of X. But in that case it does not depend on X, and thus we can remove it from $\mathbf{N}$ and add it to the witness $\mathbf{W}$ without consequence. Therefore there always exists a choice of witness so that $\mathbf{N} = \emptyset$, and thus the result follows. □

Proposition 7 *Out of all definitions we have considered,* **Def 10** *and* **Def 3** *are the only ones which do not satisfy* **Dependence**.

Proof For the HP definitions this is proven in Halpern [8, p. 26].

Example 17 shows the result for **Def 3**.

Example 15 shows the result for **Def 10**.

Therefore it remains to be shown that **Dependence** implies **Def 2**, **Def 4**, and **Def 8**. This is a direct consequence of the fact that **Dependence** implies **Modified HP**, combined with Proposition 14. □

Appendix D: Def 2, Def 4, and Def 8, vs the HP Definitions

Proposition 8 *If* **Modified HP** *with* $\mathbf{X}$ *a singleton, then* **Def 2**, **Def 4**, *and* **Def 8**.

Proof Recall the root variables $\mathbf{R}$ from Observation 1. Note that for any setting $\mathbf{r} \in \mathcal{R}(\mathbf{R})$, for any set $\mathbf{Y} \subseteq (\mathcal{V} - \mathbf{R})$, there exists some $\mathbf{y}$ so that $\mathbf{R} = \mathbf{r}$ is both weakly, actually weakly, and strongly, sufficient for $\mathbf{Y} = \mathbf{y}$.

Assume $X = x$ causes $Y = y$ according to **Modified HP** with witness $\mathbf{W}$. This means there exists a x' so that $(M, \mathbf{u}) \models [X \leftarrow x', \mathbf{W} \leftarrow \mathbf{w}^*]Y \neq y$. Let $\mathbf{S} = \mathbf{R} - (\mathbf{W} \cup \{X\})$.

First we focus on **Def 4**. Note that $(X = x, \mathbf{S} = \mathbf{s}^*, \mathbf{W} = \mathbf{w}^*)$ is weakly sufficient for $Y = y$. Furthermore, changing X from x to x' obviously has no effect on any of the values in $\mathbf{R}$. Therefore $(M, \mathbf{u}) \models [X \leftarrow x', \mathbf{W} \leftarrow \mathbf{w}^*]\mathbf{S} = \mathbf{s}^*$, and thus we get that $(M, \mathbf{u}) \models [X \leftarrow x', \mathbf{W} \leftarrow \mathbf{w}^*, \mathbf{S} \leftarrow \mathbf{s}^*]Y \neq y$. (Also, we may assume that

$\mathbf{W} \cap \mathbf{R} = \emptyset$.) From this it follows that $(X = x', \mathbf{S} = \mathbf{s}^*, \mathbf{W} = \mathbf{w}^*)$ is not weakly sufficient for $Y = y$. So taking $(\mathbf{S} = \mathbf{s}^*, \mathbf{W} = \mathbf{w}^*)$ as witness gives the desired result.

Second we focus on **Def 2** (from which **Def 8** follows due to Theorem 4). Combining the previous statement about $(X = x', \mathbf{S} = \mathbf{s}^*, \mathbf{W} = \mathbf{w}^*)$ with Proposition 2 it follows immediately that there does not exist any network $\mathbf{N}$ so that $(X = x', \mathbf{S} = \mathbf{s}^*, \mathbf{W} = \mathbf{w}^*)$ is strongly sufficient for $Y = y$ along $\mathbf{N}$.

Clearly there exists some $\mathbf{N}$ so that $\mathbf{R} = \mathbf{r}^*$ is strongly sufficient for $Y = y$ along $\mathbf{N}$. (We can start by picking parents $\mathbf{A}$ of $Y = y$ such that $\mathbf{A} = \mathbf{a}^*$ is directly sufficient for $Y = y$. Then we can take parents of all elements in $\mathbf{A}$, to get a set $\mathbf{B}$ so that $\mathbf{B} = \mathbf{b}^*$ is directly sufficient for $\mathbf{A} = \mathbf{a}^*$, etc.) But then also $(X = x, \mathbf{S} = \mathbf{s}^*, \mathbf{W} = \mathbf{w}^*)$ is strongly sufficient for $Y = y$ along $\mathbf{N}$, from which the result follows. □

Acknowledgements Many thanks to Joe Halpern, Naftali Weinberger, and an anonymous reviewer for helpful comments on earlier versions of this paper. This research was made possible by funding from the Alexander von Humboldt Foundation.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of Interests The author declares that he has no conflict of interest.

References

1. Beckers, S. (2021). The counterfactual NESS definition of causation. In *Proceedings of the 35th AAAI conference on artificial intelligence*.
2. Beckers, S., & Vennekens, J. (2017). The transitivity and asymmetry of actual causation. *Ergo*, *4*(1), 1–27.
3. Beckers, S., & Vennekens, J. (2018). A principled approach to defining actual causation. *Synthese*, *195*(2), 835–862.
4. Glymour, C., Danks, D., Glymour, B., Eberhardt, F., Ramsey, J., Scheines, R., Spirtes, P., Teng, C.M., Zhang, J. (2010). Actual causation: a stone soup essay. *Synthese*, *2*, 169–192.
5. Hall, N. (2004). Two concepts of causation. In Collins, J., Hall, N., Paul, L.A. (Eds.) *Causation and counterfactuals* (pp. 225–276): The MIT Press.
6. Hall, N. (2007). Structural equations and causation. *Philosophical Studies*, *132*(1), 109–136.
7. Halpern, J.Y. (2015). A modification of the halpern-pearl definition of causality. In *Proceedings of the 24th IJCAI* (pp. 3022–3033): AAAI Press.
8. Halpern, J.Y. (2016). *Actual causality*. Cambridge: MIT Press.
9. Halpern, J.Y., & Pearl, J. (2001). Causes and explanations: a structural-model approach. Part I: causes. In *Proc. 17th Conference on Uncertainty in Artificial Intelligence (UAI 2001)* (pp. 194–202).
10. Halpern, J.Y., & Pearl, J. (2005). Causes and explanations: a structural-model approach. Part I: causes. *The British Journal for the Philosophy of Science*, *56*(4), 843–87.

11. Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *Journal of Philosophy*, *98*, 273–299.
12. Hitchcock, C. (2007). Prevention, preemption, and the principle of sufficient reason. *The Philosophical Review*, *116*(4), 495–532.
13. Mackie, J. (1965). Causes and conditions. *American Philosophical Quarterly*, *2*(4), 261–264.
14. McDermott, M. (1995). Redundant causation. *The British Journal for the Philosophy of Science*, *46*(4), 523–544.
15. Pearl, J. (1998). On the definition of actual cause. Tech. rep., Department of Computer Science, University of California, Los Angeles, R-259.
16. Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge: Cambridge University Press.
17. Pearl, J. (2009). *Causality: models, reasoning, and inference*, 2nd edn. Cambridge: Cambridge University Press.
18. Rosenberg, I., & Glymour, C. (2018). *Review of Joseph Halpern, actual causality*. BJPS Review of Books.
19. Schaffer, J. (2000). Trumping preemption. *Journal of Philosophy*, *97*(4), 165–181.
20. Weslake, B. (2015). A partial theory of actual causation. *The British Journal for the Philosophy of Science forthcoming*.
21. Woodward, J. (2003). *Making things happen: a theory of causal explanation*. Oxford University Press.
22. Wright, R.W. (1988). Causation, responsibility, risk, probability, naked statistics, and proof: pruning the bramble bush by clarifying the concepts. *Iowa Law Review*, *73*, 1001–1077.
23. Wright, R.W. (2011). The NESS account of natural causation: a response to criticisms. In Goldberg, R. (Ed.) *Perspectives on causation*: Hart Publishing.